# Sculpting the Space of Actions

*Explaining Human Action by Integrating Intentions and Mechanisms*

Machiel Keestra

# Sculpting the Space of Actions

*Explaining Human Action by Integrating Intentions and Mechanisms*

Machiel Keestra

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

# Sculpting the Space of Actions

## Explaining Human Action by Integrating Intentions and Mechanisms

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan

de Universiteit van Amsterdam,

op gezag van de Rector Magnificus

prof. dr. D.C. van den Boom

ten overstaan van een door het college voor promoties

ingestelde commissie,

in het openbaar te verdedigen in de Aula der Universiteit

op woensdag 15 januari 2014 te 13:00 uur

door

**MACHIEL KEESTRA**

geboren te Groningen

Promotoren:
prof. dr. M.J.B. Stokhof
prof. dr. M. van Lambalgen

Overige leden promotiecommissie.
dr. J.H. Anderson
prof. dr. N.H. Frijda
prof. dr. H.J. Honing
prof. dr. K.R. Ridderinkhof
prof. dr. M.V.P. Slors
prof. dr. W. Wimsatt

Faculteit der Geesteswetenschappen

*To my mother and sister –*
*having been sculptors in their own way*

## ACKNOWLEDGMENTS

The earliest origins of opera probably lie in the ancient Greek dithyrambos, a choral song in honor of Dionysus. This ancient vocal genre, a precursor to the genre of tragedy, was initially performed without vocal soloists and it took several centuries before a singer would step forward from the chorus to deliver a solo part. Some three millennia later, operas are rather dominated by individual roles and the vocal soloists performing them and there is little room for the choir. This history testifies to an increasing emphasis on subjectivity and individuality but at the same time overlooks the fact that individual subjects develop and emerge only through interaction with their fellows. Moreover, most people learn to sing in a group, perhaps at first in a family context, later in school classes or clubs and many also perhaps in a proper choir.

In comparison with the world of vocal music, in academia the role of the choir, to continue the metaphor, has always been much less visible, or rather, audible. The choir from which individuals' thoughts emerge is implicitly present mostly in a list of references, hidden at the end of articles and books. The exception to this rule is the preface of an academic dissertation, in which an author more explicitly acknowledges the fact that his voice can only be heard thanks to the inspiration, efforts, support, responses and protests of many others. I am glad to use this prelude in that vein and mention the other members of the many choirs in which I participated and that helped me to develop my voice and vocal part.

Before doing so, however, I would like to thank both my esteemed conductors or vocal coaches, who have helped me to develop my voice and song. Being an 'external' promovendus, I realize how extraordinarily lucky I have been in having been supervised by two great experts, who have devoted a lot of thought, time and attention to this work in progress over many years. Martin Stokhof has nourished the project from its early, embryonic stage and with his continuous attention and trust has enabled it to overcome several difficult passages. Both on the macro-level of the project as a whole and on the micro-level of sentences, he has demonstrated a most welcome combination of liberalism and precision. Michiel van Lambalgen stepped in somewhat later and presented challenges to the project that were not easy to meet but have eventually made the argument much stronger and the text more accessible. I realize that I may not always have been as good a listener to them as they might have wanted and I want to thank Martin and Michiel for their patience during this period of supervision. I will certainly miss our many exchanges of emails from one M to the other two M's.

Fellow members of the choir are, as mentioned, very important. They sing different parts and by doing so give the individual singer the joy of participating in a complex piece of music that he would be unable to perform by himself. In my case, I have enjoyed the many discussions with colleagues and friends from very different choirs, academic and non-academic. Here, I would like to begin by thanking warmly those colleagues and friends – listed alphabetically – who were willing to discuss with me components of this project: Jan-Bas Bollen, Stephen Cowley (co-author of two articles which laid the foundation for chapter I.2), Nico Frijda, Nel van den Haak, Joke Hermsen, Wolfram Hinzen, Charles Hupperts, Victor Kal, Max van den Linden, Huib Looren de Jong, Harro Maas, Stephan Schleim and Lourens Waldorp. They have all contributed to this work with their own sound and voice, even though we may not always have sung unisono.

I have also enjoyed being part of other ensembles that have inspired me and made the writing of this piece possible, some of which deserve mentioning here. First and foremost, I owe many thanks to my colleagues from the Institute for Interdisciplinary Studies of the University of Amsterdam who have, in one way or another, contributed to it. Ranging from Bernard Kruithof - who was always a welcome dialogue partner about many topics including this 'academic humiliation' -, via our secretaries – who have photocopied many texts for me -, to my teaching fellows – who have often challenged me with their fresh remarks -, and literally all other colleagues: the IIS has always provided a welcome change from the soloist exercise of preparing a dissertation. In particular, I would like to mention and thank the three consecutive directors who all have supported this project in one way or another: André Schram, Jeanine Meerburg and Lucy Wenting.

This project being interdisciplinary, I gladly mention another ensemble in which I have enjoyed participating for more than seven years: the Association for Interdisciplinary Studies (formerly: Association for Integrative Studies). During the AIS conferences and via email, I have had the pleasure of engaging in dialogues with colleagues from a wide range of disciplines and institutions, sometimes about parts of this project. My board membership of the AIS has allowed me to have continuous conversations with very engaged colleagues who have always been willing to share their insights and experiences, not just concerning conceptual issues but also about nitty-gritty details of academic and other parts of life elsewhere on the globe. In particular, I would like to mention AIS's co-founder, executive director and past president Bill Newell, who has been both very welcoming and inspiring from the very first time we

met, and AIS's influential past president Julie Thompson Klein, with whom intellectual and personal exchanges seem to blend so naturally and pleasantly.

Interdisciplinarity and transdisciplinarity by their very nature perhaps attract persons with great openness, curiosity and willingness to join voices. Indeed, many thanks to my colleagues from the Philosophy of/as Interdisciplinarity Network and those from the International Network for Interdisciplinarity and Transdisciplinarity – and particularly the fellow members of the latter's Steering group – for providing room for highly divergent and exciting ensembles, which have been inspiring and enhanced my expertise in many ways.

Listing all these much appreciated colleagues reveals that preparing and rehearsing this score has kept me from taking part in several other ensembles, even though I have enjoyed them greatly for decades. Singing with friends – literally or metaphorically – is probably one of the most rewarding activities and one of the greatest downsides of preparing this book has been its taking up so much of my time and preventing me from continuing these ensembles. I'd like to reassure my kind and patient friends that their presence and our dialogues have sculpted its contents in many ways and that I'm looking forward to us joining voices again.

Naturally, my initial voice culture took place within family circles, both wide and narrow. These circles were and are occupied by very outspoken and highly different or even opposing voices, providing a both challenging and inviting environment for aspiring singers. In this context, my parents had clear and distinct voices, alternating between unisono, harmony and the occasional discord. Yet my sister Myra, brother Ruben - to whom I am grateful for the wonderful cover design - and me were always encouraged to articulate our own parts while simultaneously listening and trying to understand the voices of others. For this encouragement and their stable support over the years, I owe a lot of thanks to my parents. It is to my mother and sister, sadly lacking from our vocal ensemble which they have sculpted in many ways, that this book is dedicated.

Out of sight from colleagues, friends and even family members, each vocalist has to rehearse his part endlessly, making irritating and shrieking noises or keeping annoyingly silent. As the writing of a dissertation is a Wagnerian task, I owe much gratitude to my children, Amos and Sarai, for bearing with me during the many years of preparation. Not only were they forced to be my audience, I was less available

for our close-harmony singing than we all would have liked. Apart from thanking them for their forbearance, I'd like to acknowledge that their voices have influenced my intonation in many ways. It has been a special gift to have had increasingly the opportunity to even rehearse with them some of the lines, particularly with my colleague in spe Amos.

My final chord here is devoted to my partner Mercedes, who deserves my heartfelt thanks. She has supported in many ways over the years my protracted rehearsing times even though it affected our duo singing. Moreover, notwithstanding differences in our interests and our musical tastes, she has generously provided the basso continuo which allowed me to develop my part. Finally, being the socially engaged person she is, she has at times sculpted my vocal patterns in other, valuable directions. I hope to show indeed that the exercises from which this book is a result can also bear fruit on other, non-academic, stages.

# Contents

# Contents

## Part I – Conceiving and explaining: an intricate relation

Being the first of three parts, Part I concerns methodology. Explaining someone's cognitive and behavioral performance requires showing how the brain makes it possible, while internal and external sources of information play an important role. Such an explanation can only adequately be developed by first defining the phenomenon that is to be explained. Consequently, an explanatory method is proposed that facilitates an interdisciplinary investigation, integrating insights from various empirical sciences and philosophical analysis.

## 1 Introduction: a common capability with divergent results

Human agents are capable of learning a wide range of actions, some of which require a lot of expertise, like performing an opera role, while other actions can be carried out impromptu. Besides, some actions require explicit attention and conscious coordination, while others are realized more automatically. This book explains how we can understand and explain the fact that an expert's automatic actions, too, can still be considered intentional and subject to the coordination and organization of his actions. From Aristotle onward, philosophers and scientists have had an interest in explaining how an individual agent's behavioral repetoire, or "space of actions", is sculpted via a plurality of processes and how this is visible in his actions. 'Sculpting a space of actions' implies increasing the differentiation between habituated versus unpractised actions, between preferred versus avoided actions, further increasing the consistence and coherence between his actions. Explicit instruction by teachers, individual deliberate practice, endless repetition of motor actions and more all contribute to this sculpting process which affects behavioral, cognitive and neural processes. Given the number and diversity of these determining factors and the additional complexity of their interactions, it is evident that a complex explanation is required. Part I scrutinizes four different explanatory models, looking for a model that can cope with the diversity of determining factors and also account for the dynamics involved in the process of an agent's sculpting the space of his actions.

## 2 Concepts as delineations for empirical content

Explaining a complex and dynamic phenomenon like the process of an agent's sculpting the space of his actions requires some integration of conceptual insights with factual, empirical evidence. According to the model about to be discussed in this chapter a strict and logical distinction between concepts and facts should be recognized, in which concepts require a philosophical analysis whereas facts are the results of scientific investigation. Importantly, this account contends that it is possible

to provide a consistent and clear conceptual framework of psychological functions based upon the analysis of concept use and the behavioral criteria that are commonly used to ascribe someone a particular function. Empirical scientists who do not understand nor comply to such a framework are accordingly liable to utter nonsense when presenting their findings as evidence concerning psychological functions. However, our critical discussion points out several weaknesses in the assumptions behind this account and suggests instead that conceptual ambiguities and divergences can be exploited as heuristics for empirical investigations and that a more pluralist approach to the relation between concepts and facts should be allowed.

The influential approach to cognitive neuroscience developed by Marr explicitly acknowledges that explanations are the result of a plurality of insights which are different in kind. The account involves three different levels of analysis or levels of explanation called computational, algorithmic and (neural) implementation levels. These offer different explanatory perspectives on a particular cognitive task, which are only loosely interdependent according to Marr. The computational level concerns the goal and functionality of the task, including the logic of it in light of its wider context. The algorithmic level is devoted to options for the representation of information used in the task and its transformation during task performances. Investigation of the implementation of the task, like in the brain or in a computer, can subsequently help to constrain the options for the algorithmic level and vice versa: some algorithms would be better served with a particular implementation than others. Although this account is explicit in its acceptance of explanatory pluralism, it leaves undetermined how the different explanatory perspectives on a cognitive function can be integrated and result in a more comprehensive explanation. Moreover, the approach allegedly has difficulties with complex and interactive systems which are subject to multiple influences - such as the kind of systems that allow agents to become experts in singing and moral actions.

Another explanatory model is employed in consciousness research. Since consciousness is a complex phenomenon and lacking agreement about a conceptual framework for it, research of consciousness focuses largely on two types of research. On the one hand it focuses on the investigation of small transitions from a particular content into consciousness or out of it, while on the other hand it focuses on the investigation of different background states of consciousness like coma, sleep and conscious states.

Scientists then look for specific neural activations that can be correlated to such transitions or states: neural correlates of consciousness. Modestly avoiding a strict conceptual definition or task description of consciousness, they hope that from assembling many of its neural correlates a general account of consciousness emerges - possibly a functional explanation that might offer a substitute for a definition, or an overlapping neural process. However, two problems remain: firstly, without any preliminary definition of consciousness it is impossible to accept or reject a finding as being a neural correlate of consciousness and secondly, upon accepting a neural correlate we still don't possess an explanation of how it contributes to consciousness. For that we need to specify the explanatory mechanism.

## 5  Mechanistic explanation and the integration of insights

A model that recognizes differences between partial explanatory accounts yet provides resources for their integration, is mechanistic explanation - which is not the same as classical mechanicist thought. Mechanistic explanation explicitly requires as a first step the preliminary definition or delineation of a phenomenon, such as for example a cognitive task. Secondly, the cognitive task must be - if only tentatively - decomposed into component tasks. In some cases, a component task can further be subdivided in even smaller tasks, as has successfully been done with vision or memory. A third and final heuristic that researchers carry out is the localization of the respective tasks somewhere in the responsible organism or system. Applying these three heuristics, researchers can uncover an explanatory mechanism that is responsible for the task, consisting of component parts and operations at different mechanism levels. These components interact in an organized fashion and  in response to both internal and external conditions. Importantly, changes regarding a cognitive task that occur during development and learning depend upon changes that affect such an explanatory mechanism. Such changes can correspond either with the recruitment of a particular new mechanism components, or with a novel configuration of components, or with the emergence of new types of interaction with the environment, or a combination of such mechanism modifications.

## 6  Concluding remarks after considering the four methodologies

## Part II – Dynamics of change and stability in cognitive mechanisms

Part II is concerned with the explanation of structural changes in an agent's behavioral and cognitive performance. Insights derived from the method of 'mechanistic

explanation' from Part I are applied to four different theories about development and learning. With these results we can go on to explain how a 'sculpting process' can have enduring effects on the mechanism responsible for changes in expert performance.

## 1 Introduction: from dynamics to stability and back again

The previous Part has ended with an analysis of how development and learning corresponds with the modification of a relevant explanatory mechanism. Part II subsequently discusses different prominent explanations for changes in an individual's cognition and behavior as a result of such modifications. It is argued that such changes further sculpt an agent's space of action, for example by expanding the agent's ability for different modes of action, like when he develops both an automatic and consciously controlled mode of performing a particular action. Such an expansion is part of what distinguishes an expert, who is better able to exert control on his modes of action, from a novice. More specifically, it is argued that in many cases we can observe the formation of a 'kludge', or an extra component that is 'cobbled-together', in a mechanism, which can explain the stability of effects of such changes. Seven characteristics of such kludges are discussed, to which we will return when we will consider different explanations of the consequences of learning and development in terms of kludge formation, affecting an explanatory mechanism.

## 2 Modularization as a process corresponding to learning and cognitive development

According to neuroconstructivist accounts, development and learning correspond with increasing modularization of underlying neural processes. This modularization can be observed in both the proceduralization of a skill - it becoming more stable and flexible while also increasingly automatized and implicit - and its subsequent explicitation - rendering the skill eventually accessible for explicit correction or transformation. These developments are not just constituted by changes in the neural processes but also by the changes in the representations involved as Representational Redescription occurs. This neuroconstructivist theory of modularization can be understood largely in terms of kludge formation, affecting the explanatory mechanism underlying a skill in several ways. In other words, kludge formation can be observed in the agent's actions and involves changes in the representation of information involved as well as modification of the underlying mechanism, confirming the methodological insights presented in Part I.

### 3  Dual-process theories and a competition between forms of processing

Development and learning produce differences in an agent's performance and its underlying mechanism. Yet in many cases, the result is that an agent has more than just a single mode of performing an action at his disposal. Prominent in cognitive sciences, dual-process theories contend that an agent can perform many cognitive and behavioral tasks not just by a single type of processing but by two different types: automatic and controlled processing. A task's underlying process can become automatized after some time, which affects important properties of the task performance as it no longer requires the involvement of conscious control, nor explicit representation of the task and is usually faster. This shift from controlled to automatic processing can be partly described in terms of kludge formation jn which the underlying mechanism is modified, associated with changes in the representations involved.  An agent might employ several available strategies for controlling the type of processing by shifting his attention, by preliminary activating a behavioral schema, or by other ways. Such controlled self-regulation can itself become automatized, which can be partly understood in terms of kludge formation as well. Kludge formation thus sculpts the agent's space of actions, contributing to the varieties in his performances.

### 4  The brain as a mechanism capable of kludge formation and open to external information

We've learnt that an agent's performance can rely on several modes of processing, to which development and learning contribute as these result in kludge formation. In this third account of learning and development we will apply this notion to the emergence of a 'simulator', which essentially consist of a complex representation that is composed of components and is not stored as a whole but in a distributed manner across the brain. Such a simulator can be involved in various cognitive processes, facilitating an agent's perception, imagination and action concerning that representation, for example. Such a representation allows for recomposition or redescription and for the inclusion of component representations of environmental objects, like tools, or information, like language. Due to this, an object or word can also activate or compose a simulator in an agent, which in turn produces the 're-enactment' of previous states of perception, motor activity or cognition, affecting in many ways his subsequent performace. The notion of a simulator largely matches with our notion of kludge, like it having effects on task performance, its being composable from previously established components, its integrating environmental information. In addition, establishing such simulators further contributes to the agent's sculpted space of actions.

## 5  Dynamic mental mechanisms, kludge formation and establishing constraints on the space of options

While focusing on multiple processes like child development and skill acquisition, the shift from conscious to automatic control, learning to use tools and language and the like, Part II has confirmed how mechanisms responsible for particular functions can become modified during those processes. Such mechanism modification often implies kludge formation which has been shown to involve multiple characteristics. This process has consequences for the agent's performances as it contributes to his 'sculpted space of actions': this space can become to include novel actions while excluding others, some action options will become easily activated while others do not, the relations between actions and particular environmental conditions might change, and so on. As a result of this sculpted space, an agent's action performance can acquire a certain stability and consistence, even if many of his actions are performed through another type of processing than consciously controlled action. In addition, as an agent's expertise to a large extent depends on his having established a sculpted space of actions, this space is also involved in his capability to adjust his actions in ways that a novice can not do.

## Part III  Sculpting the Space of Actions with Intentions and Mechanisms

Part III is devoted to the explanation of expert action, such action being in many senses different from novice action. Differences between experts and novices are explained in terms of their action intentions, which are elaborated far more, but also in terms of underlying cognitive and brain processes. It is explained how experts, enabled by the process of 'sculpting the space of actions', can perform increasingly complex actions while coordinating and organizing these much better than novices.

## 1  Introduction: multiple mechanisms yet stable patterns

The same action can be produced by multiple non-identical mechanisms and mechanistic explanation helps us to account for the changes in an agent's action performance and control that occur as he gains expertise. An explanatory mechanism can become modified when kludges emerge, often by a reconfiguration of previously present mechanism components and while somehow integrating environmental objects or information. In section III.1.1. we clarify what 'sculpting the space of actions' entails and how it contributes to an agent's complex but relatively stable 'sculpted space of actions' and its internal structure. These insights can help to analyze and explain from both a philosophical and a cognitive neuroscientific perspective how different types

of intention are involved in the agent's actions and in his action planning. Indeed, we will consider the presence of a hierarchical 'intentional cascade', consisting of motor, proximal and distal intentions with each type having specific properties, involving specific representations and neural implementations. The three types of intentions play a particular role in an agent's intentional actions and also interact in several ways. Scrutinizing this intentional cascade will learn us a lot about the sculpted space of actions that an agent has.

## 2 Motor intentions: the first step in the hierarchy, or not? 226

Motor intentions are held responsible for the implicit guidance and adjustment of intentional motor actions, which are distinct from mere motor reflexes. Other than reflexes motor intentions rely on action representations that contain information about motor movements and relevant environmental conditions, which agents are capable of storing in memory. These representations are in non-conceptual form yet they are structured and are modifiable. Expertise consists in part in learning to compress or chunk motor representations and to gather many of those chunked motor representations, thus sculpting a space of actions and enhancing consistency between actions. Like simulators, these representations influence multiple cognitive processes, enabling an expert to recognize and respond faster and more flexible to relevant environmental information than a novice can. The mechanism modification at neural levels associated with expertise is twofold: at first the neural processes become more efficient, then they can become associated with other processes, allowing an expert more complexity in his actions and also leaving room for taking his other intentions into account.

## 3 Proximal intentions: a mediating role 257

Proximal intentions mediate between the distal intentions that contain representations of future actions in a conceptual format and the motor intentions that guide motor movements in response to environmental conditions. A proximal intention is responsible for quickly anchoring or perhaps instead inhibiting a distal intention in a concrete perceived situation, specifying the necessary representation components, partly based upon an agent's stored motor representations. An expert can rely upon his having assembled many relevant and complex representations, or action schemas, enabling him to usually act more quickly and adequately than a novice but also to exert more control over his actions. A mechanistic explanation of these effects of acquired expertise involves a dual process theory, consisting of an automatic 'contention scheduling' process that can be to some extent modulated by controlled supervisory

processes, each relying on distinct neural mechanism components that can interact. Expertise then implies the agent's familiarity with these processes, providing him with several options for determining and constraining his proximal intentions in multiple ways that allow him to let these fulfill their mediating role optimally.

## 4  Distal intentions: governing the intentional cascade? 297

On top of the hierarchical 'intentional cascade' are distal intentions that are in a propositional format and more abstract, requiring further anchoring and specification for their future execution by lower levels of intention. They are held responsible for governing and coordinating an agent's actions and help him to foster consistency in his actions by taking his wider web of intentions and his future actions into account instead of focusing just on a single intention in a specific situation. In order to be effective, distal intentions must be able to multifariously interact with his proximal and motor intentions and influence his sculpted space of actions, too. Agents typically do so by a narrative simulation of future situations or a more comprehensive narrative self-account. Such a narrative simulation consists of complex representations of action at several levels of hierarchy, employing the previously mentioned simulators that are stored in a distributed way across the brain and reconfiguring these in novel ways. An expert has learnt how to determine and employ his distal intentions more effectively than a novice, also by including specific schemas and cultural ingredients. Distal intentions exert their influence on future actions partly via modulation of the agent's neural 'default mode network' which has rich connections to neural networks involved in cognitive, affective and mentalizing tasks that are relevant for determining his future actions. Though not without limitations, distal intentions can thus influence these actions in several ways as they become entrenched in that network. This completes the rich interactions between the agent's intentional cascade and his sculpted space of actions and enables him to indeed become an expert - an expert who causes so often surprises and invites complex interdisciplinary explanation.

# INTRODUCTION: INTENTIONAL ACTION AND A SCULPTED SPACE OF ACTIONS

This dissertation will explore the explanation of intentional action and more in particular intentional action as carried out by an experienced agent or an expert. Imagine how an expert opera singer performing Don Giovanni is able to join in with his melody even when his Zerlina partly fails her line, avoiding to show his annoyance, to dance a short choreography simultaneously, to wink inconspicuously at Leporello, to attend to the conductor who is surprising him with novel tempo indications – behaving all the while as a somewhat ironical womanizer according to the stage directions. Clearly, such a singer does not have the time or the ability to carefully reflect and consider all these components of his performance. How is it possible that an agent – whether an opera singer on-stage or a citizen in the public domain – can act spontaneously, adequately and in line with his intentions and preliminary deliberations while also responding to unexpected events around him, without constantly interrupting his actions to reflect and decide on each successive step?[1]

This example presents us with a paradox, as it appears to pertain to expert action. Usually, we expect an expert to act fast, appropriately and adequately without pausing for considering his action options, evaluating these and finally determining the action to be executed. However, his not pausing seems to suggest that expert action is not an intentional action, as it is not the actual outcome of a consciously deliberated choice. The conclusion would then be that an expert like our opera singer does not perform intentional actions on stage. In contrast, the novice who continuously and consciously considers and determines each single component action separately, as he cannot rely on his expertise and experience, would be performing intentional actions - even though they amount to a bad opera performance. Consequently, that novice would gradually lose his capability of intentional action as soon as he starts to rely on his gathered expertise. The aim of this dissertation is to solve this paradox of expert action by offering new insights into the nature and explanation of intentional action. While doing so, we will discuss intriguing similarities between the excellence that one may reach in artistic performance and in moral action.

---

[1] Wherever reference is made to a subject or agent in this book, both sexes are obviously implied although we will refer exclusively to the male or neutral in order to avoid politically correct, gender neutral constructions. Moreover, we have chosen to illustrate our arguments with an expert singer and opera roles are usually not gender neutral.

A concept that will be introduced to achieve our goal is that of 'sculpting the space of actions' leading to a 'sculpted space of actions'. We will argue that the dynamic process of sculpting the space of actions can lead to relatively stable results, as when an expert's performance is facilitated by his sculpted action space. It should be acknowledged right away that it is not new in cognitive neuroscience to explain a certain task with the help of a multidimensional space. Several proposals have been developed that use such a notion, one of the earliest being Munsell's color sphere (Munsell 1912). Researchers have proposed to represent all colors at particular places in a space with particular dimensions and structure, as this allowed them to partly explain some peculiarities of color perception. This representational organization, mainly but not exclusively due to the photochemical properties of retinal cells and the physical properties of light, has been invoked to explain cases like the presence of an after-image with opposing color and the influence of contrast on color perception (Isaac 2009). Extending the notion to moral action, a multidimensional space has been proposed, in which moral significance and praiseworthiness are the main dimensions according to which actions are conceptualized. It was stipulated that such a space represents the results of a neural network that is trained to identify and discriminate moral actions (Churchland 1998). Extending it even further, the notion of multidimensional 'concept spaces' has been elaborated for the representation of particular informational domains that are employed by an agent's various cognitive functions in parallel and respond dynamically to his ongoing activities and situational contexts (Gärdenfors 2004b). In sum, the multidimensional, spatial representation of colors, actions or other contents has proven to be valid and useful in the interpretation and explanation of several cognitive and behavioral processes.

'Sculpting', in turn, has elsewhere been articulated as a process that influences an agent's responses, as was observable in the study of a language processing task. It was found that the 'response space' available to a subject for filling in the blanks in a sentence could be 'sculpted' or constrained by providing grammatical or other constraints on the number of answer options earlier in a text. With such 'sculpting' based upon a combination of the presented information and the subjects' previous language expertise, subjects had an easier task in determining an appropriate response, as was evident from response time and correctness as well as from the amount of neural activation during the task (Frith 2000). We will contend here that a comparable process exists with regard to intentional action, that can help to explain the paradoxical properties of expert action, a process which we will refer to as 'sculpting the space of actions'.

In this dissertation, it will be argued that intentional action rests upon a complex

process carried out by a complex and dynamic cognitive mechanism, responding to many internal and external factors, which employs a 'sculpted space of actions'.[2] This cognitive mechanism itself consists of several component mechanisms, carrying out component tasks that contribute to the performance of intentional action like perception, emotion, intention, and motor action. Many different factors determine this comprehensive mechanism, some of which have an enduring influence on it, while other factors can influence it in a more momentary fashion. It will be argued that intentions – implicit and explicit – are among the factors that can have an enduring influence. More generally, the processes of development and learning leave an important structural trace on the mechanism responsible for an agent's action. As a result, the action performed by an agent with experience in a domain of action is usually not a random response to an unanticipated situation. So instead of all potential action options having an equal chance of being determined as a response, an expert's sculpted space of actions consists of a set of action options with differential probabilities that are dependent upon long-term and short-term influences.

To show the important contribution that this process of sculpting the space of actions can make to explaining intentional action, this dissertation will cover three quite different topics. First, we will engage with the method of explanation in cognitive neuroscience. Next, we will investigate processes of development and learning while using the method of mechanistic explanation. Finally, we will use the insights derived from these preparatory parts in a discussion of both philosophical and cognitive neuroscientific studies of intentional action. It is in that last part that the notion of 'sculpting the space of actions' helps to understand some characteristics of intentional action that otherwise appear to defy explanation.

Although we will refer for the greater part to recent studies and publications, our discussion is in fact partly motivated by a long-standing debate on expert action. Before presenting an overview of the dissertation, let us provide a sketch of this background.

---

[2] The type of mechanistic explanation discussed in this dissertation is essentially different from mechanicist explanations offered by Descartes, Newton and others and has been developed over the past decades particularly as a valuable method in the life and cognitive sciences. Extensive treatments can be found in (Bechtel 2008 ; Bechtel and Richardson 1993 ; Craver 2007 ; Wimsatt 2007). Explicit discussion of the application of this model to the explanation of human action like we are doing here, however, has been very limited to date.

[3] The study of music as practiced in ancient Greece involves a much wider domain than in modern times, for example by including the performance of theatrical and epic texts, assuming music to have medical and moral value in creating balance in a person's body and soul and considering it to be a topic for mathematical and philosophical studies (cf. (West 1992)).

## How Aristotle avoids the paradox of the expert by accepting causal pluralism

The ancient debate between Aristotle and Plato on moral action provides an early example of how the paradox of expert action can be handled differently. Interestingly, in that context Aristotle also notes the similarity between musical and moral performance.[3] In Plato's dialogue 'The Republic', Socrates famously describes how philosopher-kings are exclusively capable of rationally determining their moral actions, with all others being limited in this regard. Music may offer some help in preparing the city's guardians for prescribed moral habits, because it is pre-eminently able to influence someone's feelings by providing mere imitations of real actions. Yet it is only of limited value, as it is crucially distinct from the essentially rational skills that philosopher-kings must learn in order to decide rationally about the – moral – goals of the polis (Republic, book VI-VII).[4] Musical activities and reasoning, so we are told, rely on distinct capabilities and have different effects, with the former being much less relevant than the latter. As a result, there is hardly any interaction between the two, rendering their comparison hardly useful.

Aristotle's position is very different in several respects. Rejecting the idea of a rational deduction of moral actions from a supreme and single good, he criticizes Socrates because the latter allegedly "thought all the excellences to be kinds of knowledge" and only "inquired what excellence is, not how or from what it arises" (Ethica Eudemia 1216 b 6-10).[5] In contrast, Aristotle held that regarding excellence "not to know what it is, but to know out of what it arises is most precious" (Eth. Eud. 1216 b 20-21). This critique amounts to at least two different points: first, excellences might well differ in their nature, second, the source and process of developing excellence are important to know. With the recognition of these points, Aristotle is able to avoid the paradox of expert action.

Reason is not rejected as a determining factor of expert action, yet it is robbed of its exclusivity, being only one determining factor alongside several others, to which it is often related. For example, in the Rhetorics Aristotle mentions as much as seven factors that co-determine action: "Thus every action must be due to one or other of seven causes: chance, nature, compulsion, habit, reasoning, anger, or appetite" (Rhet. 1369 a 5-6). These factors are quite different from each other, even though some are

---

[4] Musical performance as discussed by Plato generally includes text and therefore 'logos', because it concerns songs and choral parts (cf. Republic 398 C). It is with regard to the specifically musical components that strong reservations are made.
[5] Unless stated otherwise, all references to Aristotle's works are to the revised Oxford translation, edited by Barnes (Aristotle 1984).

related, as we will learn below. In the case of excellent or expert action, the number of determining factors seems to have decreased, for it is argued that "there are three things which make men good and excellent ('agathos kai spoudaios'); these are nature, habit, and reason" ('physis, ethos, logos' - Pol. 1332 a 38-39). Apparently, among other things expertise amounts to gaining some control over factors like chance, compulsion, anger and appetite. Leaving aside for a moment the developmental process involved, the message from these remarks is that Aristotle upholds a causal pluralism with regard to action.[6]

With regard to moral action, Aristotle explicitly rejected the attempt to ground moral action in purely rational decision-making, as its implications were not acceptable to him (cf. Ethica Nicomacheia III, 5). One of those implications would amount to the paradox noted above, which would apply to our opera singer as well as to a moral agent: if it is only through rational decision-making that good performances can be made, does that imply that an expert performer or seasoned and brave citizen deserves less praise from us than a novice? Is it less of an accomplishment if the expert acts in a seemingly natural way without apparent conscious efforts, reliant as he is upon previous reflections and practice, which have instilled in him several acquired habits and dispositions? Should we praise the novice instead, even though he must continuously pause to consider his actions, explicitly remember earlier exercises, reason about the consistency of his performance, meanwhile losing sight of his fickle environment? Aristotle clearly denounces such conclusions, which would render intentional expert performance a. Part of his strategy to avoid these implications of the paradox of expert action is to convince us of a causal pluralism involved in it.[7] In addition, he underlines the importance of developmental processes.

## The importance of development and learning in Aristotle's account

Let us return once more to the musical domain to articulate why development and learning are so important for the explanation of expert action and for solving its

---

[6] Aristotle's philosophy of action has received separate attention only relatively recently. Publications of Charles and Sorabji have helped to develop this domain, both offering comparisons between Aristotle's and contemporary – analytic – approaches. Both publications demonstrate the strength of Aristotle's approach in its being embedded in a more comprehensive systematic philosophical position than its successors are (Charles 1984 ; Sorabji 1980). Given his systematic interest in empirical knowledge, it may not be surprising that Aristotle's analysis of intentional action lends itself well for both ethical reflection and naturalistic explanation.

[7] Causal pluralism is especially hard to avoid in the life sciences in general because organisms are subject to a large variety of determining factors. Generally, it is associated with a theoretical pluralism as well (Mitchell 2002). Since cognitive neuroscience is part of these life sciences, causal and theoretical pluralism reign in its domain, too.

apparent paradox. Having ourselves practiced and performed a couple of opera roles – like Aeneas (Dido and Aeneas), patron Uberto (La serva padrone) and Don Jose (Carmen) - as an amateur, our admiration for accomplished opera singers has only grown. It is especially the stacking and integration of all mastered component tasks of such performances that is hard for a novice or amateur to reach: memorizing large amounts of foreign texts, innumerous notes, harmonizing with an accompanist or ensemble, responding to a conductor's baton, impersonating a character as interpreted together with a stage director, interacting plausibly with other personas on stage. Given the necessary amounts of education, practice and reflection, it is hard to believe that all of this can still result in a convincing, spontaneous and emotionally arousing performance. Nonetheless, for an accomplished opera singer, to perform a new and difficult role like Saint François might be as challenging as it is for an amateur singer to perform just a single aria from Aeneas or for a novice to sing a birthday song.[8] Apparently, as humans we are capable of gradually familiarizing ourselves with actions or action features, seemingly performing them without requiring our attention or reflection or conscious decision-making, even though they originally did depend upon such capabilities. What this capability shows us is that with increasing expertise, an agent's actions are determined by different factors.

Aristotle acknowledges the importance of development and learning in many different contexts, including the context of intentional action. During the process of learning to judge and act morally, the interaction between reason and other determining factors of these capabilities is important. Increasing interaction can be observed, for example, when someone's character determines how things or goals appear to him (Ethica Nicomacheia 1114 a 9 ff.). Interestingly and in contrast to Plato, music can positively contribute to such interaction and thus support the development of an agent's excellence. These benefits of engaging with music are not only available for the less rationally inclined agents, but even for those who are capable of the highest form of a reflective life. For music does not only have an impact on the non-rational parts of our soul but on the rational part as well. Indeed, it contributes to the listeners' acquiring the "power of forming right judgements, and of taking delight in good dispositions and noble actions" (Politics, 1340 a 16-19).[9] Apparently, both musical and moral actions invoke a whole range of capabilities that must somehow be integrated with each other. Furthermore, both rely upon processes

---

[8] It must be noted, though, that our accomplished opera singer should be taken as an ideal-type in a Weberian sense. Other than instrumentalists, singers tend to have less insight in their score, for example.
[9] Depew fleshes out how Aristotle conceives of the relation between music and contemplation, both playing an important role in his ideal of a flourishing polis (Depew 1991).

of development that are partly even continuous with each other. This makes their combined investigation for us all the more relevant.

Most significantly, musical and moral actions as performed by experts are partly determined by habits and dispositions, which are lacking in novices according to Aristotle. These habits and dispositions are the result of extended periods of education, exercise and deliberate practice, even though we may not immediately recognize this. Indeed, the fact that exercise and habituation play a role in learning to do virtuous acts is reason for Aristotle to deny that excellences come naturally ('physei') (Ethica Nicomacheia 1103 a 21). The paradox of expert action rests to a large extent on ignoring this fact.

Nonetheless, although they don't come naturally and need a lot of consideration and attention, Aristotle contends that virtuous habits can still become 'like nature' ('tei physei eoiken' - Eth. Nic. 1152 a 30-31). One sign of the naturalness of such learned activities and moral habits is the fact that they are quite enduring and provide pleasure to the agent (Eth. Nic. II, 3). In a way, then, the acquisition of habits and dispositions builds upon an agent's natural capabilities, like his capability of experiencing pleasure and pain under certain conditions. Education and practice aim then to shape or sculpt this natural capability of experiencing pleasure in such a way that an agent feels pleasure when performing a certain action appropriately and pain when he performs awkwardly.

The lesson from the preceding sections is that it is not just the plurality of determining factors of expert action that is accepted by Aristotle, it is especially their interaction during development and learning that is underlined in his analysis. Over time and with sufficient diligence, the influences of these factors merge in such a way that an agent can be said to have developed a 'second nature' – even if Aristotle didn't use that word (McDowell 1994).[10] It is with this (implicit) idea of a second nature that Aristotle aims to avoid Plato's position of only recognizing someone as performing good actions if he has explicitly and rationally chosen them.[11] The result is a complex account that challenges both our existing conceptualization and explanation of such action performances.

---

[10] Forman correctly notes that McDowell tends to overlook the distinction that Aristotle upholds between someone's second, acquired nature and his original natural state. For example, Aristotle accepts that habits remains easier to change than someone's nature (Forman 2008).

[11] According to Aristotle, one way in which such an interaction occurs is in the establishment of a 'hexis prohairetike' – an agent's state that concerns his deliberate choice and which is characterized by his desires being in accordance with his standing practice of making such choices (Eth. Nic. 1106 b 36). State and character are important notions in Aristotle's account and allow him to develop a moral psychology that is richer than Plato's, enabling him to steer free from the paradox mentioned earlier, cf. (Sherman 1989).

## Conceptual innovation in Aristotle's account of the development of expertise

Phenomenologically sound as Aristotle's analysis of virtuous and skilled action may be, it does raise several questions that have proven hard to answer or to push aside. These questions have among other things to do with the nature of intentional action and with attempts at explaining intentional action. The fact that intentional action appears to be a moving target seriously complicates both issues. First, how can we determine what an intentional action is or whether a particular behavior should be recognized as such, if novice and expert actions are distinguishable in so many respects? Even with regard to their being intentional actions, we may be able to note relevant differences, since an expert may have performed an action – sung his Don Giovanni canzonetta – rather automatically, yet still be better capable of explaining afterwards how and why he adjusted his dynamics and vocal timbre than his novice colleague might be. So the expert may rely initially more on his implicit, automatic expertise while on a second note be very able to articulate and explicate his performance, while conversely the novice may plan in detail to perform his aria in a particular way but be unable to explain why it went so embarrassingly wrong. Second, given these and other differences between intentional actions as performed by experts and novices, should we even try to offer explanations of these actions? Or should we acknowledge that intentional actions are to be subdivided between those performed by experts and others performed by novices? If we were to do so, however, we would also need to account for the fact that experts inevitably started out as novices, gradually gathering expertise of a certain domain of action. Apart from the fact that experts and novices alike would contend to engage in intentional action, creating two classes of intentional actions would conflict with their being developmentally connected with each other.

Aristotle, being a philosopher with an extraordinary interest in phenomena from living organisms, appreciated the conceptual and explanatory difficulties that stem from such issues as discussed above. Indeed, the introduction of the concept pair 'dynamis' and 'energeia', or potentiality and actuality, was meant to avoid the paradoxes that easily surface in the context of such phenomena. Indeed, in his discussion of 'becoming' he gives an example that merits mentioning here: "We can say the man becomes musical, or what is not-musical becomes musical, or the not-musical man becomes a musical man" (Physics, 189 b 35 – 190 a 1). Firm in avoiding the unsatisfactory solution that Plato – characterisable with his predilection for the immutable domain of mathematics - proposed when understanding dynamical phenomena, Aristotle developed such innovative concepts, offering new perspectives

and articulating previously implicit features of these phenomena. A seed somehow potentially contains the tree that grows out of it, a child has the potential to become a fully developed citizen, and a novice can after time become an actual expert in a certain skill: two categories of living beings that appear in many ways different from each other are then conceptually united in a novel sense.[12] Although this helps us to escape from some of the difficulties mentioned earlier, other challenges lie just around the corner.

In particular, by placing novice and expert actions on a continuum and within the same class of phenomena, their explanation becomes more complex. Instead of producing separate explanations for two distinct classes of actions, the explanation of intentional action must now be such that it can cover a range of actions. If the explanation of novice action relied on ingredients that are distinct from those included in the explanation of expert action, it would be rather difficult to demonstrate how the latter might have developed from the former. Conversely, we shouldn't expect that the explanatory ingredients of the two are identical. For example, returning to Aristotle's 'habit' as one of several determining factors of action, we shouldn't expect it to be as important for novice action as it will become for expert action, given the latter's expertise that provides such habits in the first place. Apart from being able to rely upon his habits, an expert is also capable of intentionally modifying his habitual performance in subtle ways, by carefully articulating what features he would like to adjust in the future – his tone of voice, his posture, his taking the hand of Zerlina. So even though he started out as a novice, after a long while an expert can switch between different modes of performing his action that employ the resources that are available to him in different ways and yield actions with different properties, even though we agree to treat them as belonging to the same class.

This short discussion of the paradox of expert action and of Aristotle's approach to it has yielded several insights. Before introducing the reader more specifically to this dissertation, let us review the main ones. To begin with, as our comparison of artistic and moral action and the comparison of novice and expert performance have shown,

---

[12] Although this is not the place to discuss the revolutionary nature and importance of this conceptual innovation of Aristotle, it should be noted that it washas also been an important source of inspiration for Hegel's (often misunderstood) approach to conceptual logic – which wasis in part meant to account for developmental and historical phenomena. Going into Hegel's own explicit statements of admiration for Aristotle, Hartmann's lecture of 1923 washas been influential in further analyzing this important connection, emphasizing the relevance of this conceptual innovation (Hartmann 1957 [1923]). Similarly, Ricoeur argues for the importance of the Aristotelian notion of 'capacity' in his discussions with neuroscientist Changeux (Changeux and Ricoeur 2000). For a more general appreciation of Aristotle as an empirically minded philosopher, see (Lloyd 1982) and our (Keestra 2000).

it is quite difficult to define the concepts involved in and required for the study of such complex phenomena. Second, a phenomenon like an expert's opera performance is determined by a variety of factors. This implies that any account has to provide room for a causal pluralism, usually associated with a theoretical pluralism. Third, and related to the previous insights, the dynamic aspect of the process of expertise acquisition adds another challenge. It not only implies that the phenomenon under scrutiny behaves like a moving target, it also requires that we develop and employ the necessary resources for understanding and explaining precisely these dynamic aspects of it, how the changes occur. Indeed, we have introduced and elaborated on the concept 'sculpting the space of actions' which we argued is a new and useful resource for understanding and explaining human action.

All in all, we will need an interdisciplinary investigation of expert action and offer the resources to integrate the different disciplinary insights pertaining to it. This dissertation intends to present both the necessary insights and the resources for their integration and will do do in the following three parts.

## Part I: Investigating four methods of explanation in cognitive neuroscience

Part I of this study offers an investigation of four different methodologies that are used for the investigation and explanation of cognition and behavior in cognitive neuroscience and argues that the method of mechanistic explanation is best suited for our goal. Each of the explanatory methodologies suggests a way to connect a conceptual delineation of a particular psychological function with the ingredients that we usually employ in its explanation: cognitive processes, representations of the information that is processed, and neural activities underlying these processes. However, the methodologies differ starkly from each other in how they conceive of this connection between concepts and explanations.

The first methodology we discuss - as presented in the book "Philosophical Foundations of Neuroscience" (Bennett and Hacker 2003) – defends a strict separation of conceptual analysis from empirical research, claiming that it is possible to agree upon neatly distinguished concepts for psychological functions and that empirical scientists merely have to find novel facts about those functions. We argue that the idea of such a consensus is unwarranted and defend a more complex interaction between conceptual and empirical investigations, offering a better chance for the explanation of complex phenomena.

A more complex interaction between explanatory ingredients is included in the second approach, David Marr's conception of computational cognitive neuroscience.

It breaks the comprehensive task of cognitive neuroscientific explanations down into three distinct tasks, implying the need for three different theories, each offering a different perspective on the function at stake: a 'computational theory' that describes the task that is carried out (and is an alternative for a conceptual analysis of the task), an 'algorithmic theory' aimed at clarifying the representations of information and their transformation as used in the task, and finally, a 'neural implementation theory' that considers the neural components that may be responsible for the function (Marr 1982). Although these theories can be developed relatively independently from each other, a result regarding one theory can often be used to constrain the options available for the other two, thus enabling scientists to contribute to each other's work. Useful as this subdivision of the explanatory task is, this methodology is hampered by some of its assumptions and has difficulties with the complexity and dynamics of a task like our topic: the determination of intentional action.

The third methodology to be discussed, which is used in the search for neural correlates of the complex and multifaceted phenomenon of consciousness, takes the relation between its delineation and explanatory facts for it to be looser (Chalmers 2000). Indeed, it is even suggested that the dispute over a definition of consciousness might be solved by the discovery of a neural correlate that is shared by different phenomena in the domain of consciousness (Lamme 2006). Although such a solution to conceptual problems appears tempting, we will discuss some objections that show how these problems nonetheless affect empirical results. Apparently, looking for mere correlates of a phenomenon that lacks a clear delineation may at most be useful as a first step in developing a cognitive neuroscientific explanation for it.

Part I closes with the exposition of the fourth method, so-called 'mechanistic explanation'. Recognizing the difficulty of explaining cognition and behavior, which are often hard to define and are characterized by causal pluralism, it offers several heuristics that together can help scientists to develop an increasingly comprehensive and detailed explanation in terms of a mechanism that is responsible for the phenomenon under investigation (Bechtel and Richardson 1993). A definition of the phenomenon should be followed by its decomposition in terms of the sub-tasks that appear through their interaction and in response to environmental conditions to produce it. Each of those sub-tasks in turn is somehow carried out by explanatory mechanisms that can be localized in an organism at several levels, for instance at the level of synaptic processes or neural network activations. The mechanistic explanatory approach explicitly provides the resources to integrate insights from many different disciplines and allows mutual constraints between them (Craver 2007). In addition, we point out how this approach offers the resources to explain

dynamical changes such as they occur during the acquisition of expertise, in terms of modifications of the 'explanatory mechanism' responsible for the agent's performance. Notwithstanding some limitations, we will conclude that mechanistic explanation is the most promising method for the current topic, given that intentional action is characterized by causal pluralism and developmental aspects.

## Part II: Dynamic cognitive mechanisms and their stable adjustments due to development and learning

A complex and modifiable explanatory mechanism can be invoked if we aim to account for dynamic changes that happen during the developmental trajectory along which an infant learns to speak, to sing, and then to grow into an expert singer, or more generally for the variability of cognition and behavior. After the methodological Part I, we turn to the central topic of this study, which is the complex and dynamic processes involved in forms of intentional action by an expert or experienced agent. Part II focuses on aspects of development and learning, taking up three theories that employ the resources that we will find to be included in the mechanistic explanation in Part I. These three theories are neuroconstructivism, dual-process theory and a particular simulation theory, which will be clarified below.

Part II starts by discussing more generally why and how a dynamic mechanism acquires an increasingly complex structure, in which new components develop from already existing components. These new components become stably and generatively entrenched in the mechanism and thus influence its future development (Wimsatt 1986). Such new components can be considered as being 'cobbled together' and can accordingly be called 'kludges' (Clark 1987) – a term that we will use for components established during development and learning which subsequently affect the results of the associated explanatory mechanism's activities. We will present several characteristics of such 'kludges' that are relevant for the investigation and explanation of a certain function as it changes over time. A kludge is usually observable in changes of an agent's performance and is associated with a difference in the cognitive processes involved in that function. As for a kludge's neural implementation – the domain of Marr's third perspective – there are usually several options available. Given the complexity and dynamics of the mechanism in which a kludge emerges, we should expect variability in performance between the stages of individual development and between individual subjects, as is the case when we observe performances of the same opera aria by different singers. This expectation is examined with regard to three different cognitive neuroscientific theories.

The first theory is neuroconstructivism, which connects child development

and learning in general with the increasing complexity of the brain's networks, in which modular structures emerge. The initial formulation of neuroconstructivism distinguished between two processes that can be observed in child development and take place largely in parallel: the proceduralization of a skill which renders it more automatic but less accessible for control and articulation, and the explicitation of the skill which, step by step, offers a child new ways of memorizing, adjusting and correcting its performance (Karmiloff-Smith 1992). The process of 'representational redescription' – an algorithmic theory in Marr's terms – that corresponds with learning, eventually yields several representations of one and the same task for an agent. These representations, together with their underlying modified mechanisms correspond to the kludge formation that we use to account for some of the variability observable in performances of that task.

The second theory – or set of theories - testifies even in its name, 'dual-process theory', that different yet related explanatory mechanisms must be invoked to explain the differences in action performance by humans. For aside from the controlled processing with which human agents can carry out a task, experienced agents often also perform in an automatic fashion, allowing a fast and flexible performance even of complex tasks. Such automatized actions employ representations that are stored in memory and rely upon distinct neural processes that implicitly use these, together presenting another example of kludge formation. The use of these action representations easily escapes control, which raises questions about the intentionality of actions performed in this manner. Yet even though some theories suggest otherwise, the controlled and automatic processes are related to each other in several ways. To begin with, the acquired action representations can to a limited extent be influenced – or sculpted – by explicit, conscious control during learning. Even after an agent has reached the stage where he commonly performs a task through automatic processing, there are means available for self-regulating his automatic task performance. For instance, he can use his explicit understanding of the representation involved in a task and attend to specific features of the action representation that will adjust his eventual performance. Indeed, we also recognize an expert's performance by his adequate shifting between these two modes of processing.

The third theory of learning and development to be discussed also points to kludge formation, underlining the influence of cultural symbols, language or artefacts on the representations processed during cognition and action. Observation of expert singer performances, which are partly determined by these kludges, demonstrates how such represented contents, too, can become integrated in the operations of a complex and dynamical system like the brain. According to Barsalou's simulation

theory, such representations are not memorized as a whole but its components are distributed across the brain, waiting to be employed again. Kludge formation here refers to the emergence of a specific simulator, which is a 'distributed multi-modal system' that flexibly draws together component representations containing sensory, motor or cognitive information related to a particular content, experience or action – like performing an opera role (Barsalou 2009). The automatic or controlled activation of such a simulator enables an agent to re-enact in a rich and multi-modal sense a previous experience, to imagine a future action, and so on. This idea of a simulator containing representation components that refer to contents from the environment, concurs with the theory of extended cognition which holds that the brain can even integrate objects and tools that exist outside of the skull, in cognitive routines (Clark and Chalmers 1998). These two theories of learning again confirm an agent's capability of establishing complex and hierarchically structured representations associated with his cognitive processes and actions, in which heterogeneous elements can be integrated. Such kludges build upon each other with subsequent learning, facilitating an expert's increasingly complex performance.

Part II confirms that the explanatory mechanisms that produce our cognition and action are modifiable through learning and development. Learning and development lead to kludge formation, which involves the activation of richly structured representations and the neural activation patterns associated with these, both in an automatic or controlled way. As a result, an agent can usually learn to perform a particular, complex action in multiple ways and increasingly control and determine the mechanism that produces it. This learning process can be considered as a process of 'sculpting the space of actions', a space from which an agent's actions subsequently come forth.

## Part III: Sculpting the space of actions for the performance of intentional expert action

Part III more specifically scrutinizes how actions are determined by various types of intentions that contribute to kludge formation and thus enable expert action or an expert singer to perform in a stable, flexible and fast way his complex opera roles according to his interpretation. Earlier, we rejected the paradox of expert action, because we agreed with Aristotle that an expert should be recognized for his excellent performance even though he relies upon his acquired skills, habits, and dispositions that relieve him of continuous conscious and rational decision making during his performances. Indeed, we learned that the explanatory mechanisms that produce cognition and action are modifiable and develop kludges as a result of learning

and development, which facilitate expert performance. In addition, we learned that intentional control exerts some influence upon his performance by influencing the formation of these kludges and their subsequent activation. That gave us a first indication of how an agent's intentions can be among the multiple causes that together determine his actions. In Part III of this dissertation, we will discuss more in detail how different levels of intention can contribute to stable modifications of the explanatory mechanisms that produce an expert's performance. In doing so, we will introduce and define in section III.1.1 the notion of a 'sculpted space of actions', mentioned in a more general sense earlier.

It is important to realize that for an agent to respond with an appropriate act in any given situation is a complex task, and suggests that devising or selecting such a response must be constrained. What features of the situation are relevant to act upon, which desire might now be fulfilled, what action goals can be realized and what consequences should be avoided, what is the cost-effectiveness of one action option compared to another? Consciously and rationally deciding about all these issues and weighing many others would cost a lot of time and resources, impeding his adequate and appropriate interactions with his constantly changing environment. Fortunately, thanks to the modifiable mechanisms that underlie his actions and the generatively entrenched kludges with which these have become equipped as a result of his growing expertise, he has a 'sculpted space of actions' at his disposal. No longer are all possible response options equally likely to be performed, as his response space has become sculpted, pushing some action options to the center and others to the periphery. Similarly, within a sculpted space there are dissociations or assocations between some actions, external conditions have become integrated and many other factors have influenced it. Indeed, we will argue that this sculpting process integrates not just representational contents referring to relevant environmental conditions and the agent's expertise with these, but also representations corresponding to his long-standing intentions. An expert opera singer's sculpted space of actions is thus determined by the opera roles he has practised, his vocal and acting experience, his artistic and moral convictions and so on. Consequently, unlike his novice colleague, he is able to perform his roles fast and flexibly, while paying attention to his colleagues on stage and in the orchestra pit and the stage directors' seat and responding to them in a way that is coherent with his multi-facetted intentions. How these intentions become integrated in this sculpted space and correspondingly influence the mechanisms responsible for his actions will be discussed in the remaining part of this study.

Becoming an expert singer requires careful long-term planning, persistent

learning, a lot of deliberate practice, and increasing control of the complex musculature involved in singing and acting. The differences between all intentions involved, which figure at several levels of specificity, together with their necessary interactions will here be analyzed according to a conceptual framework presenting them as an 'intentional cascade' (Pacherie 2008). This framework allows a parallel discussion of philosophical analysis and empirical results pertaining to intentions. The philosophical analysis focuses mainly on the contents, representational format and functional role of each level of intentions and the interactions between the levels. Our discussion of the empirical, cognitive neuroscientific results will be constrained by the results of our philosophical analysis, which is being used as a heuristic for that discussion – as learned previously from the methodological Part I. Starting with the motor intentions, we will continue with the intermediate level of the proximal intentions, eventually arriving at distal intentions. Concurring with our methodological observations, each of these levels can be described in terms of the task for which they are responsible, the representations and operations involved in that task, and the neural mechanisms in which representations and operations are implemented. Furthermore, development and learning have differential effects on these levels of intentions, modifying both underlying mechanisms and the action representations, confirming the results of part II.

Contrary to what looks to many like an expert's performance of mere automatic, implicit and unconscious motor movements, philosophical analysis of action points out the crucial fact that such an agent is actively guiding his movements, as is visible in his continuously correcting and adjusting these in response to environmental changes (Frankfurt 1978). The representations involved in these motor intentions contain not just non-conceptual information about stimuli and motor responses, but must be much richer and more complex to enable the guidance an expert visibly exerts on his actions. Since such complexity would put a large burden on his resources, it is welcome that these representations are compressed or 'chunked' in an expert (Gobet and Simon 1996 ; Miller 1956). As he can rely upon thousands of such chunked representations that were stored during his long-term absorption with his art, the cognitive and neural processes are different for an expert compared to a novice. Indeed, those representations can be processed with limited neural activations, consequently allowing additional cognitive processes to occur simultaneously, which is why an expert appears to be less consumed by his own actions than a novice is.

Kludge formation modifies the mechanisms that we refer to when explaining an expert's performance, but not only with regard to motor intentions. Chapter III.3 is devoted to proximal intentions, which mediate between the implicit and situation-

specific motor intentions and the distal intentions, which are explicit, conceptual in nature and future-directed. A philosophical analysis of the roles of proximal intentions teaches us that it is more difficult than just yielding a specification of a distal intention, or anchoring it in an appropriate situation such that corresponding motor intentions can be determined. For in exceptional situations, like in an avant-garde stage direction, an expert might want to block his habitual action as a competing distal intention may need to overrule the intention that is habitually realized under particular environmental conditions (Bratman 1987). The complex task of proximal intentions can be considered the result of not just one but rather two distinct processes, one interacting with the other, similar to the dual processes discussed in chapter II, 3. Connecting the two are complex action representations that are automatically put together in response to multiple factors by a 'contention scheduling' process. This process can be modulated by a supervisory process, granting the agent some control over his automatic action if necessary (Norman and Shallice 1986). With reference to these two component mechanisms and a large set of stored action representations, all subject to change as a result of expertise, cognitive neuroscience can explain how the intricate roles of proximal intentions are realized and how these contribute to an expert's sculpted space of actions.

Although Aristotle did reject the paradox of expert action mentioned above and contended that causal pluralism should be recognized with regard to action, it is still not obvious how distal intentions can contribute to what appears like an expert's automatic actions. How is it possible that we can recognize an expert singer's moral convictions and artistic style even when he tries out a new role , and not catch him in an awkward and hammed up performance? Surprising as this may seem, it is important to realize that counterproductive actions will likely occur and produce costly incoherencies if an expert's comprehensive long-term intentions were not capable of constraining his ongoing performance. Consequently, we can expect that mechanisms are in place that are responsible for doing just that, for constraining his actions in line with his distal intentions. These distal intentions, then, perhaps not so much determine in detail an upcoming action but can rather be considered as constraints or filters that co-determine the action options available to an agent (Bratman 1987). A human agent typically integrates his multiple distal intentions in a complex and hierarchically structured narrative and shares this with other agents, adjusting it carefully when necessary. In doing so, he can employ the stable narrative configurations that are part of his cultural environment, deviating from them inevitably and sometimes at wish (Ricoeur 1992). Engaging in such narratives amounts to simulating actions similar to what was discussed before. Though the

action representations involved in these narratives are more comprehensive and stretch out further into the agent's past and future, such simulations build upon the action representations at several levels of specificity that the agent has assembled over time as part of his expertise. Indeed, evidence confirming the 'constructive episodic simulation hypothesis' confirms that there is a strong interdependence between the cognitive and neural processes for such simulation of a complex action and the processes responsible for the component representations involved in that simulation (Schacter and Addis 2007a). As a result, such action simulation not only allows an agent to consider the coherence between his intentions consciously and rationally, but in itself also influences the mechanisms that implicitly co-determine his future actions – again, contributing to the sculpted space of actions that enables an expert to act promptly and flexibly in a coherent and stable manner.

Putting the reader's perseverance to the test, this dissertation thus ends its long trajectory with a more specific investigation of the relation between on the one hand, the different levels of intentions that are the subject of philosophical analysis and on the other hand, the cognitive and neural processes involved in realizing these intentions in actions. Equipped with our preparatory methodological results and with the insights about the dynamics of mechanisms as a result from learning and development, we are able to integrate the philosophical and cognitive neuroscientific approaches to intentional action. By mutually constraining each other, these approaches help us to understand and explain the amazing and admirable properties of expert action – be it the action of an opera singer who moves us when he embraces as Saint François the leprous man, or the action of a seasoned fellow citizen who courageously and carefully manages to defuse a public strife.

# Part I

# Conceiving and explaining:
# an intricate relation

# 1 INTRODUCTION: A COMMON CAPABILITY WITH DIVERGENT RESULTS*

The extent to which humans are capable of performing highly complex and socio-culturally influenced cognition and action in a seemingly effortless way remains remarkable. Ranging from motor behavior like cycling or swimming, through the use of musical or surgical instruments, up to the concentrated efforts in social actions like acting on stage or the battlefield, humans perform these actions as if they have come completely naturally. The ease and speed with which experienced performers are capable of learning and adjusting their actions enhances this apparent naturalness. This appearance, however, conceals the efforts, attention and time that were invested before such results could be reached. A way of describing these results is to observe that humans are capable of 'sculpting their space of actions'. The limitations on the available 'space of actions' are hard to define in a general sense. Interestingly, it turns out to be malleable along several dimensions: the most obvious is that it can be expanded by including ever more types of action in it through learning and practice. A second dimension is the transformation of existing contents of the action space due to learning and practice, leading to astonishing differences in the performances of similar actions by experienced versus non-experienced individuals. Even for actions that stem from a rather common domain, the effects of the process of sculpting are clearly visible. As we will see in this dissertation, this sculpting relies on a wide variety of processes, ranging from automatization through endless repetition without much cognitive effort, to conscious control of specific components of the action.

To be sure, the phenomenon of different processes contributing to action has long received attention from scientists and philosophers. It is akin, for example, to Aristotle's reflection on our ability to habituate virtuous action in his Ethica Nicomachea and Politics. At first, Aristotle appears to determine only a single cause of action when he states that 'prohairesis' or deliberate choice is the principle of action (Ethica Nicomacheia 1139 a 31). However, he continues with the observation that there is also a final cause involved and that "desire and reasoning with a view to an end" are at the origin of the choice (Eth. Nic. 1139 a 33). In fact, Aristotle elsewhere points out that there are at least seven causes determining human action, as he lists "chance, nature, compulsion, habit, reasoning, anger, or appetite" (Rhetorics 1369 a 5-6). Similarly, the moral excellence of a person depends on such causal plurality which includes not just deliberate choice, but also nature and habit (Murphy 2002). Indeed,

---

* On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively. Figure I is particularly relevant as a represention of the main contents of section I.5.

such juxtaposition of causes has enabled man, the "animal who has the gift of speech" (Politics 1253 a 9-10), to develop not just his individual moral excellence but also to develop collectively the polis which subsequently influenced human flourishing.[13]

A plurality of causes is not specific to a distinctive and intricate phenomenon like virtuous action. In several places, Aristotle compares the effects of practicing music with the effects of learning to act virtuously and then concludes that in both cases it is the formation of dispositions that is crucial (Woerther 2008). Moreover, he states that the word for development of excellent moral functioning stems from the word that refers to habit (ethos): "ἡ δ' ἠθικὴ ἐξ ἔθους περιγίνεται" (Eth. Nic. 1103 a 17). Similarly for music, even though we depend on a natural disposition for habits to form, it is only after learning and practice that we may excel in an activity like playing the zither (Eth. Nic. 1103 a 35). In this dissertation, we will follow Aristotle and analyze some of the processes involved in human action, particularly intentional action while illustrating our analysis now and then with the example of an expert singer. Both in intentional action and in singing, we will find, a comparable plurality of causes and processes is at work. In many respects these converge in producing particular results, yet in other respects they seem to yield divergent and sometimes even counterproductive results. Since the example of musical performance does not involve difficult ethical issues associated with moral action, the complexity of explaining musical performance is easier to demonstrate. Before starting the discussion of four different methodological proposals regarding such explanatory work, which makes up this part, let us give a first impression of the complexity of explaining musical performance as it offers a glimpse of some of the issues treated in this book.

It appears that the human capability for music and particularly for singing is observable in all parts of the world and in humans from all ages and perhaps has been prominent throughout human history (Mithen 2005). Indeed, in contrast to most animals, humans have a proclivity for music perception and performance from an early age on (Honing 2009; Trehub and Hannon 2006). Where humans prefer hearing music over silence, in non-human primates the situation is quite the opposite, as they prefer silence (McDermott and Hauser 2007). Infants also show a preference for hearing their mothers sing above hearing them speak (Nakata and Trehub 2004). From as early as a couple of months, infants and caregivers engage in exchanges of vocal play that can be said to be precursory to both singing and speech, although these

---

[13] This interpenetration of the different causes marks the Aristotelian explanation for the emergence of the political community, in contrast to modern political theory as the latter usually distinguishes between the non-political nature of man and the necessity for him to avoid conflict via the imposed construction of a state (Cherry and Goerner 2006).

exchanges bear yet more similarities to singing than to speech (Papousek 1996).

Apart from this ontogenetic importance of music and singing for human infants, music may also have played an important role in the evolutionary history of humans. More specifically, it has been suggested on the basis of interdisciplinary convergence of evidence that humans have evolved from 'singing Neanderthals': combined with gestures, it were music-like vocalizations that probably made up a communication system that gave the early hominids such an advantage over other animals (Mithen 2005). Providing many benefits in terms of cognitive and behavioral development, social bonding and cooperation, the subsequent evolution of mankind occurred such that its musical abilities further evolved parallel to other abilities like speech. This co-evolution is further proof of the benefits that musical capabilities still provide, even after speech became available to humans in their history as a species or as individuals (Deacon 1997).

Notwithstanding this ancient evolutionary origin and general disposition for music and specifically for singing in humans, there are huge differences in the capabilities that individuals display – differences both in their behavior as in their cognition with respect to vocal music. Though most people can share in singing birthday songs and the like without much difficulty, their performance is poor in comparison to that of an experienced opera singer. Where the former may face difficulties when asked to join in halfway into a birthday song, to shift keys or to sing in another key than their neighbours, to keep on singing while cutting a birthday cake and the like, for opera singers fulfilling such requests is their daily bread. In comparison to non-professionals these singers are faster learners and memorizers and better performers of complex scores - which they have learnt to read and analyze - , better in combining song with other cognitive and behavioral activities, better in monitoring and adjusting their song at will, and all of this with more ease than their amateur counterparts. How are these differences in the cognitive and behavioral expertise to be explained? What causal plurality is involved in musical performance and how do the effects of this plurality converge and diverge from each other? How is an individual's space of actions sculpted by that plurality, and is there adequate coherence and consistency in this action space, or does the causal plurality confuse such an action space? For if habituation were to imply that singing is no longer fit for conscious control, it would be difficult to imagine what role formal music training could play, for example. And yet, even experienced singers continue to train, aiming to improve their vocal control and to practice new scores and refresh their command of previously studied scores.

Nonetheless, research has demonstrated that the differences between experts and amateurs or novices cannot be explained in a plausible way with reference to a single

cause only. Naturally, explanations for these differences in performance capabilities have often referred to an assumed uneven distribution of innate musical talents. However, innate talent is not the primary cause, as it is extended periods of deliberate practice more than anything else that distinguishes the two groups from each other (Ericsson, Krampe et al. 1993). Indeed, the role of practice for the development of specialist expertise has been shown in a variety of domains, ranging from sports to a unexpected domain like medicine (Charness and Tuffiash 2008). So it is not some innate nature on its own that can explain the differences, nor can practice do all the explanatory work, as we need to know what the effects of practice are.

It turns out that musical practice does have effects on the physical and neural resources that are being recruited for musical activities by experts and amateurs alike. Combination of lesion studies and neuroimaging suggests that there are correlations between the differences in expertise and the activations visible at the neural level of individuals' brains. There are some neural areas specifically devoted to musical abilities, as is evident from specific deficits in music cognition or behavior in patients with focal lesions to those areas. A review of evidence suggests that extended musical practice contributes to expertise, because it leads to distinct properties of both the neural organization and the processes responsible for the tasks involved (Peretz 2006).

As research in musical expertise and performance is becoming increasingly popular in recent years, we cannot even attempt to give a comprehensive overview of relevant issues. But the short discussion above already demonstrates that an opera singer's musical performance relies on an extended period of practice involving the interaction of natural neural resources, specific socio-cultural educational processes, patience and perseverance – first relying on parental guidance, after that hopefully on character and motivation – and finally the ability to master a score in harmony with his fellow performers. Clearly then, a plurality of causes already emerges, combining explanatory factors at the level of the brain, the individual's psychological properties and environmental influences of family and society. Explaining such a multi-causal phenomenon requires handling and integrating these different determining factors.

Indeed, causal and theoretical pluralism is becoming increasingly accepted in the life and cognitive sciences. Differentiation between levels of analysis, each retaining a relative autonomy, in biological science has allowed such explanatory pluralism to flourish.[14] Instead of aiming for a unification within a single, comprehensive theory, a more modest form of integration in a complex explanatory account is more plausible

---

[14] Wimsatt has demonstrated in various contexts that complex and dynamic systems are made up by components that have a 'relative autonomy' within such systems, as not all changes of a system have an immediate impact on the properties of all components, nor vice versa (Wimsatt 2007). In a similar vein, Beatty discusses the example of the contribution of multiple genes to a phenomenon. He refers to the 'relative significance' of distinct theories, each accounting for a particular subdomain within the larger domain of phenomena that a particular theoretical plurality aims to explain (Beatty 1997).

(Mitchell and Dietrich 2006). Within the interdisciplinary domain of cognitive neuroscience, such explanatory pluralism also enables researchers to integrate insights from different disciplinary and theoretical perspectives, accounting for multi-causal phenomena related to the brain, cognition and behavior (Looren de Jong 2002). This includes the methodological step of hypothesizing 'psycho-neural identities' which in cognitive neuroscience has been very fruitful in developing pluralist explanations, without surrendering psychology and neuroscience as distinct disciplines nor implying simple reductionism (McCauley and Bechtel 2001; see more on this in section 1.5 on mechanistic explanation). In light of such pluralism, we may expect a plurality of methods and theories involved in the explanation of a phenomenon like musical performance.

In this part, we will discuss four models of scientific explanation that have been proposed as guides to how such complex phenomena can be explained in cognitive neuroscience. In spite of their differences, they all recognize that we need to include two distinct yet complementary ingredients. First, we need to define and describe the phenomenon under scrutiny with the help of an analysis of the concepts we use when referring to it. Second, when employing empirical methods to investigate that phenomenon scientifically, relating the empirical results to the phenomenon should not be taken lightly. Especially when a particular process, which belongs to the plurality producing the phenomenon, is investigated in isolation it may not always be easy to determine its exact role. Our discussion of the four explanatory models will demonstrate that the inclusion of both conceptual and empirical ingredients is in itself not enough to avoid stark differences between the models. Let us mention here in advance just in a few words how the four models to be discussed suggest the configuration of these ingredients.

The first model, discussing philosophical foundations of neuroscience, posits a strict distinction between the conceptual analysis that philosophy provides and the empirical facts that scientists can gather. Assuming that the philosophical analysis yields a consistent and adequate definition of psychological functions like consciousness and perception of pain, empirical science as such has no contribution to offer to assist with regard to conceptual problems, according to (Bennett and Hacker 2003). The second explanatory model distinguishes between three different perspectives on a particular psychological function, like visual perception. One perspective – the computational or task theory – is devoted to the definition of the task and its goals and to its decomposition in component tasks. The other two perspectives are meant to subsequently clarify the implementation of the task: the algorithmic theory should offer potential implementations in terms of the representations and

transformations involved in carying out the task and the neural implementation theory pertains to potential implementations of the task in the brain. In this case, the three theories are 'loosely' dependent upon each other and can mutually constraint each other somewhat (Marr 1982). The third model loosens this relation between conceptual or definitory ingredients and empirical ingredients even further. It is specifically developed for the investigation of Neural Correlates of Consciousness, but is being applied to other functions as well. Here, the target is to find specific correlations between a specific instance or example of conscious experience and neural activations, for example by looking for brain activation patterns that correlate with the conscious percept of a visual bi-stable object like the Necker cube (Logothetis and Schall 1989). Instead of offering a preliminary definition of consciousness, which has proven very hard to do, some researchers even hope to avoid that task and instead discover gradually a neural correlate that turns out to underlie all different instances of consciousness (Lamme 2006). Although we will argue that this model is in vain trying to avoid conceptual problems, it does hint at the fact that conceptual and empirical insights can be used to mutually constrain each other. This is what the mechanistic explanatory approach explicitly invites scientists to do, as it aims to localize increasingly detailed explanatory mechanisms that are responsible for a particular function. Starting with heuristics that demand a preliminary definition and decomposition of a function, it acknowledges that scientists may have to revisit their initial conceptual insights when the insights in responsible mechanisms suggest to make conceptual adjustments (Bechtel and Richardson 1993). To this strong plea for the integration of insights in the model of mechanistic explanation we will add an analysis how this form of explanation offers the necessary resources for explaining dynamical changes as they happen during development and learning.

After this preparatory work, we can proceed to Part II, where our focus will be on the hierarchical yet modifiable structure of complex cognitive and neural mechanisms like those responsible for action, or for singing. For it turns out that development and learning often lead to such a hierarchical structure, and that it is this structure that is responsible for the individual's enhanced capabilities in terms of increased processing speed, stability, adaptivity and diminished recruitment of cognitive resources. Explaining a complex phenomenon therefore does not only rely on the inclusion of a causal plurality but also needs to account for the dynamical changes that can affect such a structure and its properties. This complicating factor makes the lack of unanimity regarding the preferable model of explanation for this research endeavor even more understandable.

With the help of these results concerning the explanatory nature of cognitive

neuroscientific research and concerning the structure and dynamics of complex cognitive processes we will focus in Part III on the primary object of this dissertation: human intentional action. Treating in parallel philosophical analyses and empirical results with regard to action, we will find how intentional action is similarly determined by a causal plurality that together might explain both differences and similarities between individuals. Intentions will be found to function at different levels of specificity or proximity, with all levels of intentions contributing to an agent's sculpted space of actions as intentions can dynamically affect the mechanisms underlying this space. This space itself enables him to act in a more consistent way than he would have done without such a sculpted space.

## 2  CONCEPTS AS DELINEATIONS FOR EMPIRICAL CONTENT[15]

In the previous section, a comparison was made between amateur and expert singers. As we saw, differences were partly due to training and education, which had a differential impact on neural organization of the brain of these two groups and on relevant cognitive and motor processes. Nevertheless, it seems reasonable to include both groups of singers in a cognitive neuroscientific study of singing. In other words, it does not seem sensible to separate these groups and to argue that amateur and expert singers are in fact doing something completely different when they sing – making a comparison between the two groups unacceptable. Moreover, as a determined amateur singer can usually develop into an expert singer after appropriate training, the two are distinct in a gradual sense only. Apparently, no distinction in the underlying processes is enough to dissuade us from treating amateurs and experts alike as objects for a study in singing. However, it is not always so easy to decide whether two distinguishable groups can be considered to be performing the same cognitive or behavioral task.

Sometimes it is difficult to judge if observable differences between subjects force us to split a group into two –or even more- different groups with respect to a particular task. For example, are the vocalizations of monkeys to be considered as singing and can we compare their performance and underlying processes with those of human singers? Or at what moment during child development do we accept a child to be singing and not just making vocalizations? And how about those animals most kindred to us with respect to music: the birds? It may prove difficult to deny that birds are singing, even though there are differences in human and bird song.[16] Are we therefore allowed to compare their cognitive and behavioral processes with those of human singers, or will that not inform us about human singing because of the differences between the two species?

It is such conceptual questions that motivates the methodological approach to cognitive neuroscience advocated in the joint work of neuroscientist Bennett and philosopher Hacker. With their much debated book 'Philosophical Foundations of

---

[15] The present discussion of Bennet & Hacker's work elaborates on the critical articles that were published together with Stephen Cowley. Our critical review article (Keestra and Cowley 2009) received a rather harsh response in (Hacker and Bennett 2011) which we rebutted in our (Keestra and Cowley 2011). Thanks are due to Stephen Cowley for this collaboration.

[16] Distinctions between human and bird song are often made with reference to their structural properties. Especially in the light of structural properties like syntax and recursivity, qualitative differences between human and bird song seem obvious. However, these distinction then rely on the assumption that singing is for both species a form of communication of meaning – cf. (Hauser, Chomsky et al. 2002).

Neuroscience' (Bennett and Hacker 2003) they make a strong plea to give priority to conceptual analysis of psychological functions under study and to subordinate empirical studies to the a priori concepts of such functions. Their strict distinction between the results of conceptual analysis and scientific research leaves limited room for influences of empirical research on concept definitions. Given the extremity of their position, it provides a useful starting point for our current search for a proper method to align investigations of subjects that perform comparable actions yet in remarkably different ways.

## 2.1 Concepts as 'clean instruments' for neuroscience

Being a neuroscientist and a philosopher respectively, Bennett and Hacker start their jointly written volume by declaring that: "[i]t is concerned with the conceptual foundations of cognitive neuroscience – foundations constituted by the structural relationships among the psychological concepts involved in investigations into the neural underpinnings of human cognitive, affective and volitional capacities. Investigating logical relations among concepts is a philosophical task. Guiding that investigation down pathways that will illuminate brain research is a neuroscientific one. Hence our joint venture" (Bennett and Hacker 2003 1). After they declare a strict distinction between philosophical and neuroscientific tasks and state their view that there are conceptual foundations involved in neuroscience, we learn that they were motivated by a serious dissatisfaction with neuroscientific writings with regard to these foundations. For they held "a suspicion that in some cases concepts were misconstrued, or misapplied, or stretched beyond their defining conditions of application" (Bennett and Hacker 2003 1). Apart from the question what 'defining conditions of application' imply and what role these have – to which we'll return later – the picture that emerges is that the investigation of concepts does not belong to neuroscience's tasks. On the contrary, neuroscience has to accept and correctly apply the concepts when carrying out its own task. What then is that task, if it is not in any sense involved in the investigation of concepts, or in the construction or the development of new forms of application of concepts?

As we can expect from the above, neuroscience is said to deal solely with empirical issues, as: "[i]t is its business to establish matters of fact concerning neural structures and operations" or to "explain the neural conditions that make perceptual, cognitive, cogitative, affective and volitional functions possible" (Bennett and Hacker 2003 1). Establishing facts and explaining conditions are indeed empirical scientific tasks, but still their being logically distinct from the philosophical task needs to be specified. This is done by means of a parallel: "we distinguish between the statement of a

measure and the statement of a measurement" (Bennett 2007 129) – neuroscientists taking the measure for granted and employing it in their business of measuring their objects. Clarifying the logical difference, the authors go on to say that the statement of measure is 'normative (and constitutive)', while the statement of measurement is purely 'descriptive' (Bennett 2007 130). What does this relation between statement types amount to in neuroscience?

The task of neuroscience can allegedly only take place once the philosophical task of concept analysis has already being carried out. And this task is allegedly not empirical in nature but prerequisite to it. As such, Bennett and Hacker do at times compare the relation between the two tasks with the relation between mathematics and physics, for instance when they write: "[n]onempirical propositions, whether they are propositions of logic, mathematics, or straightforward conceptual truths, can be neither confirmed nor infirmed by empirical discoveries or theories. Conceptual truths delineate the logical space within which facts are located. They determine what makes sense. Consequently facts can neither confirm nor conflict with them" (Bennett 2007 129).[17] The middle sentence captures the nature of the relation between the two tasks: first, a conceptual space must be defined in which, second, empirical facts can be placed. Without a given conceptual space, it seems, empirical facts cannot make sense at all. How would this work?

How would a cognitive neuroscientific study of a particular function like action or consciousness depend upon there being a preliminary conceptual space in which facts about that function have to find their place? Such a study often requires the issues like those mentioned earlier regarding singing to be resolved: can we compare animals and humans, is there a relevant difference between children and adults, and so on. If scientists are investigating consciousness, the authors argue that similar questions can be answered once the logical space is already determined by the apriori, conceptual truths concerning consciousness: "Philosophy is concerned with elucidating the defining features of consciousness (its a priori nature). […] Neuroscience, presupposing the concept of consciousness as given, has the task of investigating the empirical nature of consciousness […]"(Bennett and Hacker 2003 403). Obviously, neuroscience has nothing to contribute to the definitory work, on the

---

[17] Acknowledging that empirical scientists are not always happy with this division of labour and the immunity from empirical critique that it renders to philosophical analysis, the authors insist upon the non-empirical nature of it and the analogy with mathematics: "[f]or neuroscientists such as Edelman to deplore the methods of philosophers as hopelessly a priori is as misguided as it would be for physicists to deplore the methods of mathematicians as a priori" (Bennett and Hacker 2003 402, cf. pp. 7, 385). What they overlook, however, is that the allegedly non-empirical nature of mathematical theorems is itself disputed in the theory of mathematics (Crowe 1988 ; Lakatos 1976).

contrary. Elsewhere they elucidate their idea of defining an object with the example of a vixen: "an animal can be said to be a vixen if and only if it is a female fox" (Bennett 2007 ; Hacker and Bennett 2011). The example certifies that this definition of a vixen does not include biological or genetic information,[18] but instead remains within the verbal realm. However, the question is whether such a conventional or nominal[19] definition adequately captures the difficulty of defining consciousness or other psychological functions. If defining such functions is more problematic, as we believe it is, this seriously undermines this methodological proposal

Let us explain our doubts with the example of consciousness. Hacker and Bennett responded to our critique of their approach in (Keestra and Cowley 2009) with the acknowledgment of assuming the following: "We *took it for granted* that we all know how to use the word 'conscious' and its cognates –for that is all that is necessary for the clarification of *the concept of consciousness*" (Hacker & Bennett, 2011, p. 411, italics in original). Crucial here is their relying upon a 'we' that 'all know' how to use this word. That their approach deserves to be called 'anti-empirical conceptual analysis' (Sytsma, 2010) is not difficult to demonstrate in the context of consciousness. A succinct survey of philosophical accounts of consciousness shows that competent philosophers have not yet been able to settle their debates concerning consciousness and conscious states (Kriegel 2006), and the presence of heated public debates about animal consciousness and euthanasia of patients in a vegitative state confirms that a public community of competent speakers has not yet universally accepted a particular meaning of those intricate concepts.

Still, according to this proposal, the definition of the concept for a function rests not just upon a single but upon two interdependent sorts of information. First, a definition of a concept relies upon its relation to other concepts. Second, and integral to the meaning of a concept, are the criteria for the use of such a concept in this view. Let us first elucidate the role of conceptual relations. The clarification of concepts and conceptual networks that we use when describing facts is carried out in analytic

---

[18] Defining a fox and even defining femininity can be harder than is often assumed – even though most people would agree on some standard defining features of gender, whereas there may be more instances when doubt about the genus of a given cat-like animal arises.

[19] Even though the authors praise Aristotle for paving the way for their type of criticism of conceptual flaws, they did not recognize that in fact, in Aristotle's Posterior Analytics, there is a transition without strict separation from nominal to explanatory or causal definitions (Charles 2000; Demoss and Devereux 1988) – in contrast to the logical distinction made by Bennett and Hacker. In addition, for Aristotle, explanatory pluralism renders definition of biological functions and properties unlike definition of mathematical objects (Gotthelf 1997). Like Aristotle, I think that this also holds for psychological functions: these vary both in different kinds and within a single individual. Thus bodily aspects are needed in analysis, description and explanation of psychological functions (van der Eijk 1997).

philosophy in the form of a *'description of our conceptual scheme'* (Bennett and Hacker 2003 439, italics in original). This conceptual scheme is nothing new, they themselves say, it is even "the ordinary conceptual framework". Ordinary indeed, for: "[i]t consists of the familiar array of concepts we have all acquired in the course of mastering the humdrum psychological vocabulary of sensation and perception, cognition and cogitation, imagination and emotion, volition and voluntary action, which we employ in our daily lives" (Bennett and Hacker 2003 114). Nonetheless, these ordinary concepts are being compared with instruments, which tend to have a more specific function and use.

Even though the concepts making up the framework are not specifically designed by or for neuroscience, they are said to function inevitably as "spectacles through which psychological phenomena are viewed and understood." Since spectacles interfere with a person's vision, there is a risk involved: "[i]f these spectacles are askew, then neuroscientists cannot but see the phenomena awry" (Bennett and Hacker 2003 115). Apparently, even though spectacles are usually made with a specific function to fulfill or to compensate for a specific person's vision deficit, the authors hold that ordinary concepts can similarly be considered to be askew or not. Confirming this is their statement that words: "are the instruments of thought and reasoning" and their insistence that it "behoves us to be aware of our instruments and to ensure that they are clean" (Bennett and Hacker 2003 381). In sum, in spite of its being ordinary and non-scientific in nature, our conceptual scheme or framework can allegedly be analyzed – and corrected, if necessary - in such a way that it provides lay-persons and neuroscientists alike with correct spectacles or clean instruments. Even thought we doubt the appropriateness of this comparison of concepts with functional instruments, in the next section we will show where the authors believe that we find our conceptual instruments or how we can adjust our conceptual spectacles.

## 2.2 Connective analysis and ascription criteria

Cleaning our concepts, which we need as instruments, is partly carried out by a method Bennett and Hacker write about in a methodological section on 'Connective analysis in philosophy'. There they write that such a connective analysis: "traces, as far as is necessary for the purposes of clarification and for the solution or dissolution of the problems and puzzles at hand, the ramifying logico-grammatical web of connections between the problematic concept and adjacent ones" (Bennett and Hacker 2003 400). The web of connections should inform about the "logical possibilities" or the "combinations of words [that] are significant and can be used, within or without science, to say something true or false" (Bennett and Hacker 2003

401). The description of this result cannot be compared to cleaning instruments or correcting vision, as the latter activities allow gradual improvement, while logical possibility does not. Indeed, a logical possibility implies a definitive answer to a question like: "[w]hat kinds of things can be coloured – that is, what are *intelligible* subjects of colour predicates" (Bennett and Hacker 2003 130, italics in original). But this latter example is quite specific and involves the logico-grammatical relation between subject and predicate that is of particular concern to the authors and which they discuss with regard to the mereological fallacy that they find to be commonly made in cognitive neuroscience writing. We will come back to that later, but will first consider more closely how a connective analysis can deliver the 'defining features' for a psychological function like consciousness.

The nature of the connective analysis that should deliver the necessary web of connections is rendered relatively clearly at the beginning of the section on one of many forms of consciousness: transitive consciousness. "Transitive consciousness lies at the confluence of the concepts of *knowledge, realization* (i.e. one specific form that acquisition of knowledge may take), *receptivity* (as opposed to achievement) of knowledge, and *attention* caught and held, or given. The various categories or kinds of transitive consciousness that we have distinguished are differently related to these. We shall sketch some of the connecting links and some of the conceptual differences between these loose categories" (Bennett and Hacker 2003 253, italics in original). What can be learnt from this statement is that a particular form of consciousness is indeed being analyzed with the use of – 'adjacent' - concepts that are useful for describing or defining transitive consciousness. For instance, transitive consciousness can be described as a form of knowledge about an object, it being a knowledge that is not actively achieved or attained. Instead, the contents of transitive consciousness are, according to this analysis, merely being noticed, realized or one just becomes aware of them (Bennett and Hacker 2003 253). Such establishment of a conceptual framework when defining a concept does seem useful. What remains unclear, however, is what the source of the relevant web of connected concepts is and precisely how they are so sure about the relations between concepts when describing a phenomenon like this.[20] For instance, one could wonder whether previously attained knowledge influences transitive consciousness, heightening the receptivity of a

---

[20] There are places when Bennett and Hacker are less certain or where they acknowledge that strict delineations are difficult to achieve. An example is emotions. Notwithstanding the remarkable conciseness of the chapter – only 25 pages, in contrast to some 130 for (self-)consciousness – they introduce an uncommon subdivision of affections into emotions, agitations and moods, only to admit later that the: "boundaries between emotion, agitation and mood are not sharp" (Bennett and Hacker 2003 202).

subject as it does for some objects more than for others. If so, should we distinguish different forms of transitive consciousness? How should we decide such cases, where do we find the criteria to decide one way or another?

The authors do not appear to have much doubt about such matters regarding the source of the array of concepts or their applicability, as was evident from the quote above in which they referred to common knowledge about the use of the word 'conscious' and related words. Apparently, their equation of meaning and use – inspired by their interpretation of Wittgenstein - has found an uncomplicated application in the context of the conceptual foundations of neuroscience, even with reference to not undisputed concepts like consciousness. But these disputes will not easily affect the approach of Bennett and Hacker, since they put some conditions in place such that their assumption of consensus is not easily threatened.

The assumed consensus is grounded in the existence of a community of speakers, which – perhaps tacitly – has determined correct and incorrect explanations for verbal meanings. In so doing, words within such a community have a rule-governed use, which in turn determines their meaning (Bennett and Hacker 2003 382). Two additional conditions further restrict the source of word use grounding the investigated conceptual definitions when the authors state that they rely on "what competent speakers, using words correctly, do and do not say" (Bennett and Hacker 2003 400). The conditions of competence and correctness of use do to a large extent overlap or define each other reciprocally: incompetence in language use is observed especially through the incorrect use of verbal expressions, and vice versa. Combined, these conditions here depend again upon the presence of conceptual consensus within a given community. As a result, the authors modestly claim to offer only: "the ordinary conceptual framework properly elucidated" (Bennett and Hacker 2003 114), intending to be uncontroversial and merely "to outline distinctions which are familiar and in constant use" (Bennett and Hacker 2003 117).

Such consensus must be assumed as it also provides the basis for a speaker's competence to develop: "[a] competent speaker is one who has mastered the usage of the common expressions of the language" (Bennett 2007 146). They illustrate the latter with examples that refer to words black, vixen, perambulate, man and ten o'clock. Avoiding discussion here of the potential disagreements on particular instances of these words, even though we believe these are all less complicated than 'conscious', let us end this section with some more information on the criteria for use, since word use plays such an important role in this approach. Indeed, words and concepts will be found to be of importance for the other methodological approaches as well, so the present discussion prepares us for the treatment of those as well.

Closely related to the connective analysis, laying bare the conceptual framework or the web of connections between concepts, there is a second source of information about their meanings. This source is derived from observing cases in which these concepts are or are not being used. According to the authors, psychological concepts such as consciousness, perception and emotions are being used meaningfully only in relation to other human beings.[21] The rules that appear to be determining such application of these concepts are related and complementary to the rules that determine the connections mentioned above: "[t]he criterial grounds for ascribing psychological predicates to another person are conceptually connected with the psychological attribute in question. They are partly constitutive of the meaning of the predicate" (Bennett and Hacker 2003 83). To such ascription criteria of a psychological function belong particularly the behavioral expressions that are connected to that function. In the case of pain, for example, it is pain-behavior like moaning that is relevant: "Pain-behaviour is a criterion – that is, logically good evidence for being in pain", the authors write, and they conclude: "[t]hat such-and-such kinds of behaviour are criteria for the ascription of such-and-such a psychological predicate is partly constitutive of the meaning of the predicate in question" (Bennett and Hacker 2003 82).[22] They emphasize that the fact that behavioral criteria are partly constitutive of a concept's meaning distinguishes these criteria from being mere inductive evidence.

In contrast to such behavioral and non-inductive evidence, neuroscientific investigation of psychological functions like pain or consciousness, aims to produce inductive evidence concerning the brain events associated with such a function. Preliminary to such scientific investigation, the authors argue, a non-inductive and logically sound ascription of pain or consciousness can and needs to be made to a subject that is being neuroscientifically investigated. That ascription rests upon the investigators using these words correctly, including controlling whether the behavioral criteria for application of these words are being met. If a subject is

---

[21] Sytsma justly emphasizes that B&H fail to produce empirical evidence with regard to the words that they subject to connective analysis and has consequently called the method of PFN 'anti-empirical conceptual analysis'. To underline this diagnosis he produced empirical evidence that, contrary to the authors' intuitions, a significant portion of subjects don't hesitate to apply the verb 'calculate' to computers –though B&H reject this as nonsensical (Sytsma, 2010).

[22] Debate about behavioral criteria is likely to emerge, especially with regard to an elusive phenomenon like consciousness. Indeed, an fMRI and clinical study of patients diagnosed with only vegetative consciousness has shown that some patients were able to willfully change their conscious state in such a way that it was detectable with imaging techniques, in the absence of any distinct behavioral responses (Bennett and Hacker 2003 202). However, such an approach has been criticized with reference to the behavioral criteria required by Hacker, as in (Monti, Vanhaudenhuyse et al. 2010) – these criteria would still not be fulfilled with fMRI evidence. An interesting alternative has been proposed, namely to use a brain-computer interface as a way of facilitating behavior to patients without any muscular control (Nachev and Hacker 2010).

not meeting those – non-inductive, behavioral – criteria, the inductive evidence derived from the neuroscientific investigation cannot be correctly correlated to the psychological function that the investigators believe to be scrutinizing. The logical order is such that only if the ascription criteria are met, the empirical evidence can be inductively correlated to the alleged function: "if such inductive evidence conflicts with the normal criteria for the ascription of a psychological predicate, the criterial evidence overrides the inductive correlation" (Bennett and Hacker 2003 83). If applied to an example like vixen, this seems evident: if closer inspection of a particular animal that a scientist calls 'vixen' produces evidence that the animal is in fact a female wolf or that it is a male fox, further evidence about it does not apply to vixens, too. This observation does however merit further specification: evidence about gender specific features or about features that are used to distinguish wolves from foxes may no longer be applicable. Therefore, it may make sense only for such limited examples and in a limited sense to declare that: "[c]onceptual truths delineate the logical space within which facts are located" (Bennett & Hacker, 2007, p. 129). Indeed, we may doubt whether interdependence between scientific facts and conceptual truths can be avoided, for example concerning an animal's gender and its precise species definition.

Generally, natural kind concepts and classifications in the life sciences and behavioral sciences lack the kind of unity and demonstrate much more divergence than can be found in domains with less complex and dynamical phenomena, like chemistry (Dupré 2001). The reason is that phenomena studied by the life and behavioral sciences are generally produced by a much greater and wider range of causes, which simultaneously determine these phenomena. Correspondingly, any attempt at delineating a logical space that consistently and comprehensively encloses only those facts that pertain to an allegedly definable psychological function must take a variety of criteria and logical connections into account. Otherwise, the conceptual space runs the risk of resting upon ill-founded assumptions about a domain's contents, like its definability and the uniqueness of its corresponding definitions (Hacking 1991). In the final section on this approach that aims to define conceptual foundations of neuroscience, we will demonstrate where it runs into trouble and what consequences can be drawn for the relation between psychological functions, the concepts that correspond to these and neuroscientific evidence with regard to these.

## 2.3  Non-convergent and variable criteria, and their implications
In a field where causal pluralism affects relevant phenomena, exceptional and

surprising cases will likely obtain. Referring to our example of singing again, we doubt whether people would always agree in deciding whether or not a person who is making vocal sounds is singing. At what age do we ascribe 'singing' to an infant and not just babbling? Similarly, are religious recitations instances of singing, or rather peculiar intonated readings? Will we agree on when a speaker of a tonal language, like Mandarin, has shifted from speaking to singing? Are there perhaps even cases in which we ourselves unwittingly made such a shift? The blurred distinctions between speech, recitation, babbling and singing as well as our probable disagreements suggests that such concepts are in fact formed and used as prototypes, rather than definable under the conditions suggested by Bennett & Hacker.[23] Since there are conventional concepts like 'bachelor' or perhaps 'vixen' that are rule-governed, our language probably contains both prototypical and rule-governed concepts (Ashby and Ell 2001). The advantage of concepts as prototypes is that speakers can apply such concepts with some liberty and still remain understandable.[24] A strictly delineable conceptual space does not allow such liberty in use, as the disputed status of 'blind-sight' demonstrates.

This example refers to investigations of a famous patient, who was found in specific behavioral experiments to demonstrate 'good visual discrimination capacity in the absence of acknowledged experience' (Weiskrantz, 1997, p. 19) – behaviorally responding above chance to a stimulus, whereas she explicitly denied perceiving that stimulus. From this surprising combination of behavioral evidence for, yet verbal evidence against perceptual discrimination in this patient, Weiskrantz concluded that she suffered from 'blind-sight' (Weiskrantz 1997 19). This was how he addressed the issue that the facts collected in studying this patient did not permit insertion in either the conceptual space for 'visual perception' nor in the space for 'blind'. Indeed, if we consider these spaces for a moment as 'logico-grammatical' Venn diagrams, one could even imagine that the spaces 'visual perception' and 'blind' overlap at some point. In this overlapping area, then, the facts pertaining to this patient could be located. However, Bennett & Hacker argue otherwise.

---

[23] Stokhof points out that for Wittgenstein it is not just the normative practice of rule-following that constitutes the meaning of concepts. In addition to these, constraints imposed on our practices by our environment and our human nature have an impact on our conceptual schemas (Stokhof 2000), which are not articulated in Bennett & Hacker's approach.

[24] A prototype theory of psychological concepts has been proposed – on various grounds – by Paul Churchland (Churchland 1988). Elaborating on the ideas of Churchland and others, another conception of concepts as state spaces can be found in (Gärdenfors 2004a). These authors contest the assumption that concepts are always rule-governed (or symbolic). Although our approach to the process of 'sculpting the space of actions' has some affinity with theirs, we aim to show how rational considerations can also contribute to this process of sculpting a state space that pertains to actions.

Their conclusion regarding 'blind-sight' is straightforward and relies on their assumption of the strict definability of psychological concepts, partly constituted by their behavioral criteria. To begin with, they observe that in this patient "the normal convergence of indices of sight –namely, appropriate affective response, behavioural reaction, reoriented movement, verbal description, answers to appropriate questions, etc. – is subtly disrupted." Then they refer to their assumption that "such *convergences constitute the framework* within which verbs of vision are taught and used. (...) The consequence of a conflict of criteria is that one can neither say that the patient sees objects within the scotoma nor say that he does not." Finally, their conclusion from this is that this patient's case "indicates the *inapplicability of a concept* under special circumstance" (Bennett & Hacker, 2003 396, italics not in original). With concepts that function as prototypes, this conclusion of conceptual inapplicability is avoidable. Apart from the conceptual dispute, it is important to realize the consequence for the empirical evidence gathered by investigating this patient: according to Bennett & Hacker it will have little relevance for the explanation of normal vision. Before explaining this position and then contrasting it with the cumulating evidence for the divergence with regard to psychological concepts and behavior, let us underline what is at stake in the present case of blindsight.

Given their assumption that psychological concepts can be assigned strictly delineated logical spaces for which both logico-grammatical and behavioral criteria are to be used, there is principally no room for divergences regarding the use of those concepts. This also holds for those cases where some criteria for the use of a particular concept are met, while other criteria appear to be contradicted. Such divergence of criteria would allegedly render a concept meaningless and consequently useless. Accordingly, a concept is never applicable in those situations in which conflicts arise with respect to the criteria that should determine the use and hence the meaning of the concept. In contrast to a prototype theory of concepts that allows some distortion and divergence in the formation and use of concepts (Ashby and Ell 2001), as does a theory of concepts that projects a multi-dimensional state space for a concept (Gärdenfors 2004b), Bennett & Hacker cannot permit any flexibility in the criteria that constitute the meaning of concepts. Indeed, in their response to our critical review article (Keestra and Cowley 2009), they compare the correct application of concepts with following the rules in a game where those rules in fact constitute the game. This odd metaphor brings them to take on a judge-like function when writing: "Far from delimiting neuroscience or narrowing its scope, we constrain it only in the sense in which one constrains draught players in pointing out that there is no checkmate in draughts – which is no constraint" (Hacker and Bennett 2011

461). However, the arguments here and in our rebuttal (Keestra and Cowley 2011) suggest that if this metaphor of concepts as rule-governed games holds at all, it has very little value for psychological concepts. For with regard to psychological concepts we should expect, for various reasons, a variability and divergence that makes a different theory of concepts more appropriate. Such a theory could then also allow the scientific investigation of an extraordinary case like blindsight some relevance for the explanation of normal vision. Bennett & Hacker, on the other hand, cannot allow such an applicability of insights in blindsight to cases of normal visual perception.

The reason they offer to deny that a patient that we diagnose as and call 'blindsighted' can yield any neuroscientific insight on perception is as follows. Given their assumption that the behavioral criteria are partly constitutive of the concept 'seeing' or 'vision', the acceptance of the contradictory criteria that are applicable to this patient would in fact imply that we change the concept itself. Given the connections between concepts – that are subject of a connective analysis – such a conceptual change could not be made without in turn modifying all those concepts related to 'seeing' or 'vision'. Eventually, the consequences would be wide-ranging for many concepts and phrases in which these figure. When redefining a word like 'perceiving' or 'remembering', neuroscientists would be obliged to do the following: "New formation rules would have to be stipulated, the conditions for the correct application of these innovative phrases would need to be specified, and the logical consequences of their application would have to be spelled out. Of course, if this were done, the constituent words of these phrases would no longer have the same meaning as they have now. So *neuroscientists would not be investigating the neural conditions* of thinking, believing, perceiving and remembering at all, but rather those of something else, which is as yet undefined and undetermined. But this is patently not what neuroscientists wish to do" (Bennett and Hacker 2003 384, italics not in original). Or, applying once more their metaphor mentioned in the previous section, the neuroscientists that investigate 'blind-sighted' patients would play chess while those that focus on normal vision are playing draughts or even baseball – precluding any useful exchanges or competition between the two.[25]

An interesting asymmetry emerges between neuroscientific results pertaining to an exceptional case like a patient with 'blind-sight' and results pertaining to normal

---

[25] In the terms that Christensen & Sutton use in their discussion of an integrated approach to moral cognition, Bennet & Hacker would assume that it is possible to construct a 'clean taxonomy' for such a cognitive function. Christensen & Sutton, in contrast, argue that we cannot avoid 'messy taxonomies' for such functions due to the "complex underlying causal factors that overlap across categories" (Christensen and Sutton 2012).

subjects. Indeed, although the authors believe that although neuroscientists "can brilliantly explain why patients cannot behave as normal humans can in a multitude of different ways" (Bennett and Hacker 2003 365), explaining normal functioning cannot refer to such neural conditions. In contrast with explanations of pathological behavior, to "explain typical human behaviour, one must operate at the higher, irreducible level of normal descriptions of human actions and their various forms of explanation and justification in terms of reasons and motives (as well as causes)" (Bennett and Hacker 2003 365). Even though causes are added – albeit in brackets – to the list, these are not subsequently clarified like the other ingredients. So it remains unclear whether these causes refer to the tendencies or habits of an individual, to the moral and social norms or to other not explicitly mentioned ingredients. In any case, the authors appear to render only a secondary role to causal conditions when humans are explaining each other's typical behavior, even though humans typically accept that causal conditions at the neurophysiological level do offer bottom-up constraints on someone's behavior. We will come back to that in section I.2.4. Here, we would like to add another argument why we believe that the authors' assumptions are not warranted, suggesting as they do that it is always possible to make a clear distinction between normalcy and pathology and suggesting that there is always consensus concerning the use of psychological concepts within a community of competent speakers.

In contrast to these assumptions, researchers tend to accept that divergence is prevalent in the context of behavioral criteria, concept use and even neural correlates of human psychological functions. Textual analysis and interpretation have long suggested historical and cultural diversity in these contexts (Lloyd 2007 ; Snell 1975). In addition, psychological and psychiatric experiences suggest that subjects of different cultures do not only differ in the use of psychological concepts but also in their expectations of behavior corresponding to the psychological functions referred to with these concepts (Chaturvedi and Bhugra 2007).[26] Additional insights on etiology and clinical phenomena support the proposal that the strict distinction between pathological and normal states cannot be upheld, whereas a more gradual distinction between those states seems more plausible (cf. (Hyman 2007 ; Newsome, Scheibel et al. 2010)), adding to the divergences.[27] A main reason that such divergences

---

[26] Arguing for a more dynamical mode of classification, Hacking points to fact of a 'looping effect' of psychological and psychiatric classifications. Such classifications tend to influence the groups to which they apply, making people labeled as ADHD or multiple personality disorder patients behave according to the criteria currently used by a classification system like the DSM (Hacking 1995).

[27] Hyman, a member of the DSM-5 Task Force, is highly critical of the classificatory 'silos' of the current and future editions of the DSM. One of the arguments against its classification is that it corresponds poorly with clinical and scientific evidence about distinctions. He suggests integrating clusters of interrelated syndromes into larger clusters – avoiding the assumptions of strict borders between diagnoses altogether and allowing room for additional scientific insights in this context (Hyman 2011).

may have escaped notice of scientists from various fields is that the overwhelming majority of subjects used in research are drawn from a very small and specific selection of the world's population (Arnett 2008 ; Henrich, Heine et al. 2010). As emerging results of transcultural neuroscience show that transcultural differences are likely to affect not just functional networks but perhaps even anatomical structures in the brain (Han and Northoff 2008), this limitation of research subjects has serious implications for the validity and significance of its results. Such divergences are due to the long-time exposure to different cultural experiences and behavioral practices (Park and Huang 2010). Given such evidence and accounts of divergences in neural activations and structures, in behavioral experiences and criteria, and in psychological concepts, there is reason to question the consensus within a community of competent speakers, as assumed by the authors' approach. If this consensus is to be found both in concept use and regarding behavioral criteria, it may be limited to a rather restricted community. Although the authors' ambitions are larger, their conceptual foundations of neuroscience may in fact not transcend its origin as a form of "contemporary English philosophical anthropology" (Quante 2008), as a reviewer of Hacker's categorical account of human nature has elsewhere suggested. Avoiding such a serious limitation, in the next and final section on this approach, we will defend a more liberal stance with respect to conceptual and behavioral divergences, while sustaining Bennett & Hacker's critique on mereological fallacies in neuroscience.

## 2.4 Heuristic use of conceptual divergences, yet with limitations

In section I.2.1 we found that the project of developing conceptual foundations of neuroscience was mainly inspired by "a suspicion that in some cases concepts were misconstrued, or misapplied, or stretched beyond their defining conditions of application" (Bennett and Hacker 2003 1). Neuroscientists do tend to offer factual results of neuroscientific investigations as having implications for our interpretation of psychological concepts. That is, they sometimes believe they can redraw a conceptual space on the basis of those facts, instead of merely gathering facts that either belong or do not belong to a particular, predefined space. This alleged neuroscientific hubris and misapplication is warded off by presenting a 'mereological principle', which in itself is a consequence of the authors' analysis of the nature and origin of psychological concepts, as outlined above. The principle states that: "psychological predicates which apply only to human beings (or other animals) as wholes cannot intelligibly be applied to their parts, such as the brain" (Bennett and Hacker 2003 73).[28] Doing so, Bennett & Hacker argued, would be similar to chess players applying the rules for draught or bridge and thus constructing an altogether different game.

The motivation for the mereological principle depends partly on their rejection of ontological and explanatory reductionism. Scientific reductionism, they write, "is a commitment to the complete explanation of the nature and behaviour of entities of a given type in terms of the nature and behaviour of its constituents" (Bennett and Hacker 2003 357). In the case of neuroscientific analysis of human cognition and behavior, reductionism would imply that these are completely explainable in terms of neurons and neuronal activities. Bennett & Hacker have warded off this threat of reductionism by strictly separating the analysis of psychological concepts logically from the collection of empirical facts and, second, by disallowing the application of those concepts to objects other than the persons to which competent speakers ascribe them. This leaves no room for any identification of cognition and behavior with the properties, activities or interactions of neurons. However, as we will argue in this section, they overlook the possibility that other relations between psychological concepts and the study of the relevant neurons or neuronal activities are possible and even fruitful. Indeed, we will argue that a challenge for cognitive neuroscience is to develop a more useful integration of conceptual analysis and empirical research.

For Bennett & Hacker, it is straightforward that on the basis of our knowledge of the conceptual scheme of psychological concepts and of our observation of a person's behavior that we can only conclude that this person is perceiving or knowing or feeling – and not in any sense that his brain or neural areas are performing those functions. Obviously, the brain and neural areas are involved in producing a person's behavior but only in the sense of: "causally necessary conditions for the human being to think or perceive, imagine or intend" (Bennett and Hacker 2003 117). Given the nature of the sources of our conceptual truths, there is no room in this approach for these causal conditions to be more directly related to concepts like thinking, perceiving, imaging or intending– or something like their 'concept spaces'. Instead, causal conditions or correlates, being the result of empirical research, are held to be logically different and separate from those conceptual truths.

If, however, our arguments above are sound, then this strict distinction and the endeavor as a whole is flawed. If, that is, the assumption of consensus regarding

---

[28] A critique of the authors' limited account of mereological reasoning, their overlooking of the heuristic use of such reasoning and their misinterpretation of Aristotle's warnings in this context was given in (Keestra and Cowley 2009). Though largely dismissing our critique in their response, they did not address this issue (Hacker and Bennett 2011). We, in turn, reconfirmed our limited agreement with their mereological principle, albeit for different reasons (Keestra and Cowley 2011). The relativism inherent in Aristotle's analysis of part-whole relations is also commented upon in (Koslicki 2007). This relativism is better accounted for in the mechanistic explanatory approach, which is also interested in constitutive relations yet explicitly acknowledges the validity of an explanatory mechanism for a particular phenomenon.

psychological concepts within a community of competent speakers is unwarranted and if a consistent and comprehensive delineation of spaces for such concepts is illusory, then we must look for a different relation between empirical, neuroscientific facts and conceptual insights. Consequently, concepts can be considered differently and may yield insights different from those allowed by the approach of Bennett & Hacker. As we will see, the other methodological propositions that we will be discussing in this part suggest a different relation and do allow different roles for concepts and conceptual analysis. Let us finish here by discussing a few possible implications of understanding this relation differently.

If it is impossible to provide a comprehensive and consistent delineation of conceptual spaces pertaining to psychological functions, then we may need to accept and even explore the conceptual divergences and uncertainties that abound in this domain. For instance, competent language users commonly refer to the phenomenon of distraction of attention from pain. Admitting that this phenomenon defies their assurance that "there is no difference between having a sensation and feeling a sensation", Bennett & Hacker refer to this phenomenon as a "curious anomaly" which "can be viewed as a singularity (in the mathematical sense) in the grammar of sensation" (Bennett and Hacker 2003 121, footnote 2). What they fail to do, however, is to take the expression seriously – even though it blurs some alleged conceptual distinctions – and to explore its value as a heuristic. Such a heuristic use of a concept that is hard to define can point us in the direction of an explanation of its intricate character (Keestra and Cowley 2011).[29]

For instance, it may be that the causal conditions involved in pain and in attention do interfere at times with each other, producing this curious phenomenon – as was shown to be the case (Valet, Sprenger et al. 2004).[30] Indeed, when such a phenomenon is being explained with reference to a complex and dynamic explanatory mechanism – which will be clarified more generally further below in this part – its exceptional nature can be ascribed to uncommon interference of components or operations, or to external conditions that influence the explanatory mechanism such that it produces an irregular behavior. Consequently, the apparently strange concept use then

---

[29] Another type of 'bi-directional' interactions between conceptual analysis and empirical research is presented in Northoff's neurophilosophical methodology (Northoff 2004). Kindred as his approach is, it involves a particular use of philosophical analysis and pays not so much attention to the heuristic use of conceptual divergencies, for example.

[30] A similar explanation has been offered for synaesthetic experiences, which appears to correlate with cortical hyperconnectivity (Rouw and Scholte 2007). B&H pointed out that it makes no sense to ascribe colour to numbers (Bennett and Hacker 2003), which from a strictly semantic point of view may be correct but denies such a concept the role of a heuristic for further investigation of an exceptional psychological phenomenon.

correlates with the exceptional behavior of an explanatory mechanism that produces a surprising phenomenon.

In the case of 'blind-sight', a similar implication may be drawn. It is hard to define comprehensively both 'seeing' and 'being blind', even in healthy subjects, as odd perceptual phenomena occur which suggest temporary or specific forms of blindness.[31] The concept 'blind-sight' signals this blurred and porous character of conceptual definitions. Moreover, it also captures the divergence that is generally observable in the realm of psychological functions, even though we might agree in the case of 'blind-sight' that it refers to an exceptional and pathological phenomenon. Because of divergence and the corresponding ambiguity of psychological concepts in normal situations, competent speakers may at times refer to explanatory components in order to disambiguate their concepts. Given the variety of explanations, such explanatory components may be of various natures.

Aristotle, for example, aimed to define anger by including both a psychological and a physiological explanatory component in it, when he referred to anger as requiring a definition: "as a certain mode of movement of such and such a body (or part or faculty of a body) by this or that cause and for this or that end" (De Anima 403 a 27-28).[32] Filling in the required causal pluralism involved in such a definition, he specified anger as being both "a craving for retaliation" and "a surging of the blood and heat round the heart" (De Anima, 403 a 31- b 1). More generally, Aristotle accepts that such a causal pluralism is involved in human behavior and cognition, including nature (Murphy 2002). In the previous section, we adduced arguments that confirm this causal pluralism to be effective in causing divergence and corresponding conceptual ambiguities or misunderstandings. Further below, we will discuss how it is that such causal pluralism can be held responsible for the divergences that obtain in the domain of psychological functions and that transpire to the conceptual scheme when describing or explaining such functions. Instead of holding on to strict conceptual delineations that are illusory, a different handling of psychological concepts seems in order. A critical yet more tolerant conceptual analysis can indeed be more conducive to empirical research. That is, the use of the concepts themselves should be different, and the relation of the concepts to the facts derived from neuroscientific investigations can be established differently.[33] An

---

[31] Many such phenomena depend on typical perception-action loops, causing Noë to defend an enactive view of perception (Noë 2004).

[32] As we noted in (Keestra and Cowley 2009), Aristotle is not opposed to mereological reasoning, as it can perform useful functions in science. Pellegrin even argues that Aristotle's biology is in fact a mereology, a study of parts (Pellegrin 1987) – which seems to me to neglect the prominence of Aristotle's ambition to integrate the various causal contributions to a function or a kind.

influential methodological proposal which does so, has been made by computational neuroscientist David Marr. It is to this proposal that we will now turn.

---

[33] At this point we would like to refer to a comparable interdisciplinary endeavor as Bennett & Hacker's, though with a strikingly different tenor. Hermeneutic philosopher Ricoeur and neuroscientist Changeux agree, in contrast to them, that in this domain a fair amount of semantic tolerance is inevitable, if not without semantic criticism. Although acknowledging the risk that when neuroscientists employ 'semantic short-circuits' they are "illegitimately converting correlations into identifications," they do not aim to correct this with the assumption of strict delineations of conceptual spaces (Changeux and Ricoeur 2000 40).

# 3 DAVID MARR AND THE INVOLVEMENT OF CONCEPTS IN MULTI-LEVEL EXPLANATIONS

When we hear someone making vocal sounds, we still may ask whether or not that person is singing. Given variabilities between persons, generations, cultures, and so on, the behavior may need to be interpreted or defined otherwise than singing. Obviously, most aspects of the vocal expressions and causal conditions will remain the same, irrespective of our recognizing it as a case of singing. So the ambiguity in our classifying it as 'singing' may not be such that we mistake it for different functions like writing, gesturing, or other expressive actions: the voice must be involved and the sounds must be intended to having some expressiveness in order to qualify as singing. However, there are many different techniques of singing, which do not all employ our vocal tract in the same way or require the same cognitive processes of keeping melody and rhythm, or harmonizing with other instruments or voices, for example. Furthermore, whether or not vocal sounds are fulfilling a particular function is not easy to decide: we would in some cases accept vocal sounds as singing if they did not have an expressive function, for example, even though perhaps in most situations – as suggested from an evolutionary perspective – singing does fulfill a particular function. To grasp the function of singing we must usually consider the context in which the vocal sounds are being made, but in many cases we don't need information about the context to decide whether someone is singing.

In other words, there are many different sorts of information that can be brought to bear upon someone's singing, which can all be employed to decide whether that person is singing. If we want to explain that person's behavior – whether or not it is singing – we will most probably need many different kinds of information. As noted above, in the life sciences such pluralism abounds. This has led to a situation in which biologists have to recognize the only limited significance of a particular theory, allowing room for the involvement of other theories on the same phenomenon (Beatty 1997). Theoretical integration is then a legitimate explanatory goal, while unification by way of a 'single model of multiple causal factors' is not, since: "contingency, context sensitivity, and nonlinear interaction among contributing causes preclude the success of these types of unification" (Mitchell and Dietrich 2006 78). The conclusion that Mitchell & Dietrich draw is that biological phenomena allow analysis and explanation at several levels of analysis and preclude unification. Obviously, in such a situation, reduction of a complex function to a particular level is even less an option.

In the previous section we discussed the arguments of Bennett and Hacker against

a particular form of reductionism: a reductionism according to which psychological functions are directly ascribed to the brain or particular parts of the brain (Bennett and Hacker 2003). They defend their position with a particular view on the nature and origin of the concepts that we use for psychological functions. On those grounds they emphasize that neuroscientists should recognize the fundamental task of preliminary conceptual analysis for any empirical study of the functions associated with these concepts or their neural correlates. What is lacking in their particularly critical account is a proposal contributing to a more fruitful relation between such a preliminary, conceptual analysis and the empirical studies aiming to clarify the neural correlates of the investigated function. For instance, as we suggested above and in (Keestra and Cowley 2011), the results of such analysis should not just function as a barrier to nonsensical judgments, but rather provide a heuristics to suggest novel investigations. As we will see in the next chapters of this part, there are other approaches possible that agree in the rejection of reductionism and accept the mereological principle, while differing from Bennett & Hacker with their strict separation of conceptual analysis and empirical investigations. These approaches explicitly leave room for the causal and theoretical pluralism that seems appropriate for this subject. The first one that we will be discussing has had a large 'inspirational influence' in the field (Glennerster 2007) over the past decades. It may not come as a surprise that it was established by one of the authors explicitly reproached by Bennett & Hacker: David Marr.[34]

## 3.1 The analysis of computations or tasks – not concepts - should guide scientific investigations

Conceptual analysis of 'singing' can yield a description of singing as a particular type of vocal sounds, often consisting of words set to melodies, often expressing a particular intention or mood. Such an analysis does offer limited information about the function of singing itself, nor is it particularly helpful in determining scientific investigations – other than denying that scientists are in fact observing a singing person, when his behavior does not consist of expressing vocal sounds, for example. An alternative analysis would not so much analyze the concept –although that will remain an important step – but would analyze the function to which it applies by

---

[34] They cite Marr writing that "our brains must somehow be capable of *representing… information…*" and subsequently criticize such verbs as representing information, decision making and the like to the brain (Bennett and Hacker 2003 70), which suggests indeed that Marr and others have neglected the mereological principle. Marr's multi-level explanations allow a loose interdependency between such forms of analysis, precluding such mereological reasoning. Instead, Marr's approach invites the heuristic use of analytic results we defended above and earlier (Keestra and Cowley 2011).

asking questions like: what kind of a task is singing, and are there subtasks that we can distinguish? For example, during song a person is coordinating semantic and grammatical knowledge in correspondence with tonal, dynamical and rhythmic components. This involves various cognitive tasks, but also an increased demand of motor control, determining the tension of the vocal chords, breathing behavior, and so on. A task analysis that helps to distinguish these components or sub-tasks that – in coordination with some other tasks probably - make up the task of singing enables researchers to investigate such a complex task that might otherwise remain unmanageable. Notice that such a task analysis is different from, though perhaps assisted by, a conceptual analysis, even though it can equally guide and constrain empirical research of singing. Marr has contributed importantly by arguing for the relevance of such a task analysis and for describing its relation to other types of investigations in the study and explanation of vision.

Marr elaborated his methodology while working on vision or visual perception. Notwithstanding the fact that vision has traditionally been a most promising subject for cognitive neuroscience and showed quite some progress, what Marr found to be lacking in this field was a delineation of what the object of vision research in fact is. For a while, he was impressed by successes from the so-called 'feature detector' approach in vision science, primarily aimed towards discovering particular brain cells that respond to specific features in a visual scene. An example is the successful discovery of the bug-detector in the frog retina and some similar discoveries in its wake. Scientists then assumed, in the words of Barlow - quoted by Marr - that: "the activities of neurons, quite simply, are thought processes. This revolution stemmed from physiological work and makes us realize that the activity of each single neuron may play a significant role in perception. (p. 380)" (Marr 1982 13). Barlow believed that 'a single neuron could 'perform a much more complex and subtle task than had previously been thought' and Marr confesses that: "[t]ruth, I also believed, was basically neural" (ib. 14). It is such reductionist belief that was countered in the previous sections, promoting a fundamental role for conceptual analysis for cognitive neuroscience. However, Marr developed another methodological response to it.

The hypothesis that explanations of visual perception could rely on the activities of so-called 'grandmother cells' and its kin, turned out to be fruitless (Marr 1982 15).[35] For one, it turned out that the number of such feature-detecting cells that

---

[35] Apart from the fact that it is implausible to be able to find such a grandmother cell if it were a single cell in the multi-billion cells, it is also principally impossible to falsify that the cell would not respond to any other face or figure. Indeed, it is more likely that instead of single 'grandmother cells', the brain contains 'ensembles' of cells that together represent a complex object, each cell responding to a different aspect of the stimulus (Gross 2002).

were discovered was extremely limited. More importantly, even if the 'apocryphal' grandmother cell were detected, Marr realized that crucial questions would remain like: "why or even how such a thing may be constructed from the outputs of previously discovered cells" (Marr 1982 15). Since we normally expect from explanations to clarify why or how certain facts hold, this particular lack of answers in the feature-detector approach dissatisfied Marr. Trying to make up for the lacuna, he proposed a methodology that allows combining the traditional results of neurophysiological and psychophysiological investigations of vision with answers to questions about the functional role of such facts. Such answers are presented in a so-called 'computational theory'[36] and added to the other theoretical results that already figured prominently in the explanation of vision: results stemming from neurophysiological investigations and algorithmic descriptions of the neurophysiological properties and activities that were delivered by such investigations. In the next section we will discuss how Marr envisions this connection between different types of results in the explanation of a cognitive function like vision. For now, let us look more closely at what this computational theory amounts to – especially since it is this aspect of his methodology that Marr considers a particularly important contribution to the field (Marr 1982 330).

It is important to realize that a computational theory is concerned with a so-called competence[37] theory, only aiming at the formulation of the ends of a particular task, without consideration of the specific means to reach those ends (Marr 1977a).[38]As such, the computational theory focuses on theoretical ingredients that are not included in the feature-detector approach, since the latter is mainly interested in the implementation of a particular task. In contrast to Marr, Barlow and others believed to be able to deduce from evidence concerning the implementation as task its relevant functional properties, as can be seen from Barlow's influential first dogma of neuroscience research: "A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells and of a nerve cell's response to

---

[36] Marr's use of 'computational' has led to some controversy. For instance, it is different from that of Fodor, who is concerned more with 'how' veridical features of the environment are represented (Kitcher 1988).

[37] Marr himself refers to the computational theory as a competence theory (Marr 1977a), alluding to this term that Chomsky coined for the cognitive science of language, making a distinction between a competence and the way how such a competence is actually performed. As the term is here not without difficulties, I'll make only limited use of it. As it seems that Chomsky has mistaken Marr's computational theory for an exclusively internalist theory – about which more below –comparison of their ideas is beyond the scope of my present discussion. See (Silverberg 2006) for a defense of Marr's computational theory against Chomsky's interpretation of it.

[38] Elsewhere Marr made the comparison with the development of the theory of thermodynamics, which shows that a top-level theory may be useful even in the absence of : "a description in terms of mechanisms or elementary components" which appeared only afterwards (Marr and Poggio 1977 2).

such influences from other cells, is a complete enough description for functional understanding of the nervous system." (Barlow 1972, 380, cited by Marr, 1982, 13) The computational theory that Marr envisaged, on the other hand, was clearly not a functional understanding that could simply be extrapolated from properties of single nerve cells. On the contrary, Marr argues that such a method easily overestimates the relevance of an understanding based upon neural cell properties alone.

Although it is interesting to discover feature-detectors in a visual system, only a computational theory can help us to realize that such detectors might be simply misled in practical reality. For example, a 'bar-detector' may be misled, since an edge of light may be mistaken for a bar if perceived by a single cell (Marr & Hildreth, 1980, 188). Indeed, it is reflection upon such potential flaws that emphasize the relevance of a computational theory when explaining a visual system. Such a theory might force the scientist to acknowledge the importance of 'the discovery of valid constraints on the way the world is structured' (Marr 1980). What these constraints amount to will be clarified below. First, a specification of the computational theory in more abstract terms is required.

Now the analysis of the task at hand, referred to by Marr as a 'pure competence theory' (Marr 1977a) or the computational theory, does not so much concern a logico-grammatical analysis of the meaning of the concept referring to that task – like it was demanded by the approach discussed in chapter I.2. Instead, a computational theory informs us with regard to a particular task about: "What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?" (Marr 1982 25, figure 1-4).[39] Nonetheless, a first answer of those rather abstract questions may be found in the case of vision when we reflect for a moment on the question what seeing in fact is: "The plain man's answer (and Aristotle's, too) would be, to know what is where by looking. In other words, vision is the *process* of discovering from images what is present in the world, and where it is" (Marr 1982 3, italics in original). According to this plain description, the goal of vision is: getting to know something about the world, and in particular what is there and where it is. This is a rather general description, perhaps most remarkable for the fact that it describes vision without explicitly relating it to other cognitive functions

---

[39] Bechtel even argues against Marr's calling the upper level a 'computational theory': "The name is misleading since at this level the researcher is not concerned to explain the computational procedures, but rather to specify the task to be performed by the computation system, why that task is to be done and the constraints the task itself imposes on the performance of that task" (Bechtel 1994 4). It appears that Marr was at the time following the MIT convention of labelling any information processing task a computation, instead of a task or function. This convention was perhaps only superseded with the development at MIT of robots with a subsumption architecture, without a crucial role for a comprehensive representation, as in (Brooks 1991).

– a limitation that has provoked dissent, as we will see in section I.3.5 and in part II.

Leaving the latter issue aside, the computational theory still asks us questions that are more specific than the question that led to the 'plain man's answer' regarding vision. Especially the aspects of appropriateness and of the logic of the strategy used for a computation merit further discussion. What these aspects imply can be illustrated with computational theories that concern specific subtasks of vision. Examples of components of vision for which a computational theory has been developed by Marr are the recognition of shapes from contours (Marr 1977a), stereopsis (Marr and Poggio 1979), and the detection of edges (Marr and Hildreth 1980). For instance, according to its computational theory, stereo vision requires that for two separate images – one for each eye – their symbolic descriptions are being matched, where the disparity between the images is being measured (Marr and Poggio 1979). How this is to be done and what neural implementation may be responsible for it is another matter. Relevant for the computational theory are still such ingredients that are relevant for "the logic of the strategy by which it can be carried out" (Marr 1982 25, figure 1-4) and which do not refer us to the performance specifics of the system or organism under investigation.

## 3.2  Constraints that co-determine the computations' appropriateness

Marr's analysis of the computational theory appears at closer reading to consist of two components. First, it offers a description of a particular task in rather general terms. Second, it then offers information on additional ingredients that make that computational theory plausibly effective. Both components are not of a particularly technical nature. For instance, he considers the paradoxical fact that when we perceive only figures' black contours, as in Piccasso's 'Rites of Spring', these "tell us more than they should about the shape of the dark figures" (Marr 1977a). The competence to be explained is therefore how we can derive information about the objects when we are perceiving only two-dimensional black contours, sometimes even overlapping outlines of different objects. The explanation Marr offers depends upon two ingredients which necessarily underlie this capacity: "implicit in the way we interpret an occluding contour, there must lie some *a priori* assumptions that allow us to infer a shape from an outline" (Marr 1977a 441-442, italics in original) – where these assumptions concern the surface and structure of perceived objects. Such assumptions are more generally involved in cognitive tasks and should therefore be included in the computational theories underlying these tasks.

Marr uses several examples in his discussion of such assumptions. For example, the appropriateness and the logic of the computations that we perform with a cash register

in a shop rely upon "the rules we intuitively feel to be appropriate for combining the individual prices" and these rules "in fact define the mathematical operation of addition" (Marr 1982 22). The unexpected nature of such rules or assumptions stands out even more clearly when he discusses an airline reservation system that functions appropriately, depending on its capability of taking into account relevant properties of the world.[40] These properties must constrain the task if it is to provide results that are useful for the subject. Marr argues that an explanation of the computational theory of that system must not only refer to the properties of its computers, but also needs to include information about "what aircraft are and what they do; about geography, time zones, fares, exchange rates, and connections" and then expands that list even with "something about politics, diets, and the various other aspects of human nature that happen to be relevant to this particular task" (Marr 1982 5). This list of information may not yet give us much insight into a computer, nor does it yield insights in the specific ways in which the computer makes reservations, but the list is necessary to understand the kind of computations it carries out when processing reservations. Indeed, an airline reservation system that is not constrained by the presence of different time zones or local political unrest will yield results that are unrealistic or conflicting with other flights. The logic of its computations therefore depends partly upon its handling of these constraints. A feature-detector approach to the reservation system, without a computational theory of making reservations, would in this case not be able to explain why geographical differences or politics are processed by particular parts of the computer network. Consequently, it would be difficult to determine how detectors are related to environmental or internal parameters, to judge whether or not connections between detectors are functional, and so on.

One may believe that the inclusion of such constraints in a computational theory of a particular function makes the explanation of that function ever more complex. And to the extent that it forces the investigator to reflect upon a function's environmental conditions and consider which conditions are potentially relevant, this is certainly true. On the other hand, such constraints facilitate the task of explaining why a particular computation's strategy – being therefore only partly determined by its algorithmic and implementation theories, as we will see shortly – is indeed

---

[40] In this context, it has been discussed whether Marr's theory is methodologically individualistic, or rather externalist in nature. The strict requirements that Morton poses on 'non-solipsist' accounts of vision or cognition, and which Marr's approach does not meet, do not appear convincing to us. However, we must leave that topic here aside, but see (Kitcher 1988 ; Morton 1993 ; Silverberg 2006) for different positions in this discussion.

appropriate and adequate.[41] For without any constraints provided by the environment a system for vision would have much more difficulty in recognizing 'what is where': if our environment were not occupied mostly by rigid objects that have spatial contiguity, for instance, visual recognition of objects would be quite difficult (cf. Marr 1982 209).[42] Obviously, the visual system that operates in such an environment would require an explanation that includes a very different computational theory. In any case, when investigating the process of vision, researchers devote part of their efforts to discovering "additional constraints on the process that are imposed naturally and that limit the result sufficiently to allow a unique solution" (Marr 1982 104). These constraints are in fact "properties of the visible world that constrain the computational problem and make it well defined and solvable" (Poggio 1981 259). For these properties in turn determine what kind of information is needed to solve these problems and must therefore be produced by the system (Marr 1982). As a result, these constraints do not just contribute to the process of successful visual information processing, but at the same time also facilitate vision's explanation.

Compare vision with our example of singing again: when investigating the singing of a lightweight bird in the sky we will look for very different sound producing body parts than if we focus on singing whales that live in a completely different medium and have different body properties. This example also shows what seems to be lacking in Marr's approach: a consideration of the possibility that a competence may have different functions for different animals which may be partly dependent upon their other competences and even upon general properties of their existence and environments. Are the criteria for appropriateness of vision in humans not different from those in eagles or in rabbits? Even though it seems useful to focus on a particular competence in isolation from others, it may also risk leaving out important ingredients of an adequate explanation. The other approaches to be discussed in this part will take up this issue more explicitly.

Nonetheless, Marr's requiring a computational theory in the explanations offered by cognitive neuroscientists has already expanded their methodology significantly.

---

[41] Whether Marr's approach leads in fact to an optimization theory with regard to both the targeted computational problems and the strategies used for solving those, opinions diverge. See e.g. the discussion between Kitcher and Gilman (Gilman 1994 ; Kitcher 1988). It seems assumed by Marr, indeed, that evolutionary selection processes have contributed to the development of a computation characterizable by a 'logic of the strategy' (cf. Marr 1982 105, 266). Dennett argues that behind both the optimalization and evolutionary aspects of Marr's approach is his engineering stance (Dennett 1989 310-311), which seems to be a plausible hypothesis.

[42] Although Marr believes that ecological psychologist Gibson came closest to developing a computational theory, he criticized Gibson for underestimating the information processing task of vision (1982, 29). However, Marr in turn underestimated the distinctness of Gibson's approach in that it underlines the relation of vision to an animal's other competences like action (Dennett 1989 310-311).

It has forced them not just to focus on the means by which certain competences are carried out, but also to first analyze those competences and their interaction with many other conditions. The question remains, however, how this computational theory has to be related to the other results of cognitive neuroscientific research. We will address this issue by focusing on Marr's strong defense of a multi-level explanation, which has meanwhile become prevalent in the field.

## 3.3  Two further levels for multi-level explanations

With the addition of a computational theory to the formulation of explanations of vision, Marr did not aim to cast aside the more traditional explanatory ingredients. Having learnt from ecological psychology to view: "the problem of perception as that of recovering from sensory information "valid" properties of the external world" (Marr 1982 29), he continued to explain vision in terms of visual information processing as being carried out by animal brains or other devices. As a result, the computational theory has to be considered as a part of a larger explanatory framework, since explaining vision requires several distinct theories simultaneously. Generally speaking, Marr concluded that: "[f]or the subject of vision, there is no single equation or view that explains everything. Each problem has to be addressed from several points of view – as a problem in representing information, as a computation capable of deriving that representation, and as a problem in the architecture of a computer capable of carrying out both things quickly and reliably" (Marr 1982 5). The other points of view to be taken into account are presented as distinct levels, to be added to a multi-level explanation. How these levels are interrelated and how they together make up an explanation will be discussed in section I.3.4. First, we will provide short accounts of the two other points of view, or levels, in two subsections on the algorithmic and the neural implementation levels.

### *3.3.1 The algorithmic level and the representation of information*

Singing from sheet music does not come naturally. Since there are different forms of notation, it requires specific education. Depending on the instrument one plays or one's voice, within our Western tradition of music notation a musician must even learn to read in different keys and – if she plays organ - learn to read figured bass notation. Moreover, the hierarchical representation of musical information also plays a role in listening to music, as practiced listeners turn out to have different expectations than others (Justus and Bharucha 2001). Such cognitive representations develop over time and brain activation patterns together with behavioral responses show differences between individuals, ages, and even gender in the brain's musical

information processing (Koelsch, Grossmann et al. 2003). Independent of the contents of the auditory stimulus, the recognition and processing of information appear to differ widely. The representation of information has a great influence on these differences in musical competence, just as it does for vision.

Depending to a large extent on the constraints offered by the environment, the computational theory was found to focus on the 'what' and 'why' of a particular competence. The competence of perceiving objects in the world can be understood, according to Marr, as a "mapping from one kind of information to another" (Marr 1982 24). Given this information processing perspective, it is not surprising to find that: "[i]n the center is the choice of representation for the input and output and the algorithm to be used to transform one into the other" (Marr 1982 24). As the examples of recognizing objects in black contours or seeing depth from two different, two-dimensional retinal images illustrate, it is not just the gap between the environmental appearance and its perception by us that needs to be explained, but also how the information is represented in the retinal images and how their combination then allows stereopsis.

The focus of the algorithmic theory, present at the level between the computational and the neural implementation levels, is the rather abstract but crucial issue of how information gets represented and the related issue of the transformations applied to this represented information. The study of information representations and transformations is typically done in psychophysical research, employing response-time, delay, and interference tasks, or mental rotation tasks (Marr 1982 26) to develop hypotheses about "the scheme for a computation" (Marr 1980).[43] The relevance of such a scheme or representation of information is clear once one realizes how the computation of addition, for example, can be performed with Roman or with Arabic numbers or even by using one's fingers. Each of those representation forms has its own merits and disadvantages and subjects may shift from one to another representation format as the occasion demands.

What is important to note, as Marr emphasized, is that the choice of a representation format is not just a matter of usefulness. In fact, "there is a trade-off; any particular representation makes certain information explicit at the expense of information that is pushed into the background and may be quite hard to recover" (Marr 1982 21). He illustrates this not just with the comparison of Roman and Arabic numbers, but also with the necessary conversion of a human decimal number system to the binary

---

[43] Untill the work of Shepard and Metzler in 1971 visual psychologists had not recognized the importance of an algorithmic theory, according to Marr (Marr 1982 10).

representation format suitable for electronic devices. Addition and subtraction, recognition of powers of ten or equal numbers are for us much easier to carry out in a decimal system than in a binary one. Represented in a binary representation format, we would even have difficulty recognizing powers of ten, for instance. It should not be surprising, as we will indeed find out later, that differences in the implementation of computation tasks do affect the choice of the representation format. The centrality of this algorithmic level will become more evident in section I.3.4, where Marr's account of the relation between levels of explanations will be discussed.

There being several representation formats to choose from, the question arises how and on what grounds an engineer chooses one, or which format appears to have had most success in evolutionary history. Marr here makes a distinction between human vision and the visual information processes in animals. The feature-detectors that were successfully found in animals, like the frog's bug-detector and the rabbit's hawk-detector appeared to depend not just on particular hardware which may differ between animal species, but also on the particular algorithms that are performed with this hardware. A particular retinal image in combination with some other information immediately makes the animal perceive a potential prey or predator, allowing it to respond fast – even if at times it will attack or flee for an innocent object. Such representation errors are the cost that comes with the fact that each animal "can confidently be expected to use one or more representations that are nicely tailored to the owner's purposes" (Marr 1982 32).[44] Humans, however, are exceptional in this respect, Marr believes.

Unlike animals, humans do not just have 'special-purpose mechanisms' but human vision "seems to be very much more general" (Marr 1982 32). Given the wide range of information that humans can derive from their retinal images and the different computations that they perform on them, Marr argues that human vision must use representation formats – especially in the so-called early stages of visual information processing – that do not force information into the background in such a way that it is not recoverable for potential use in later processing.[45] Developing a plausible algorithmic theory for human vision should therefore obey the 'principle

---

[44] Analyzing more in detail fly vision, Marr suggests that perhaps 60% of its very simple vision representation consists of three components that allow it to detect a suitable landing spot and a potential mate (Marr 1982 32-35), admitting for the rest very sparse information about the world.

[45] Even if Marr was right about the existence of these early vision sketches, it could still be the case that they are an "an accidental by-product of its function in guiding motion for the purpose of avoiding danger and securing food, shelter, and other objects of desire" (Hatfield 1991 177). The converse would also be possible: vision may most of the times be the result of the animal's active 'probing' of its environment, but that does not preclude the possibility of an by-product in the form of a general-purpose sketch – a possibility that would counter some arguments against Marr in (Noë 2004).

of least commitment': "one should never do something that may later have to be undone" (Marr 1976 485). Feature-detectors do not comply to this principle, as their informational focus is on very particular parts of information to the detriment of other information that could otherwise be used in later processes, potentially correcting earlier misperceptions.

This emphasis on the general purpose of human vision also made Marr deny an important role for a memorized database of objects, that many researchers at the time thought would enable quick recognition in humans. Since such recognition in the early stage of vision would depend on an algorithmic theory or representation format that would foreground particular features of information, it would potentially come at some cost with respect to other information features. This is related to the supposed 'modularity' of early vision, which we will address at the end of our discussion of Marr's work. Apart from that, the assumption of the general expediency of human vision as opposed to the more specific functionality of animal vision has met with serious criticism.[46] Be that as it may, the fact remains that representation format has implications for visual and other cognitive processing. More relevant to the present context, it underscores that for a comprehensive explanation of a cognitive function like vision we need to take into account several forms of insights, stemming from different scientific disciplines and integrate these in an interdisciplinary explanation. Insights that may at times be quite disparate, presenting a challenging task for those who strive to combine them.

### 3.3.2  The implementation level and neuroscientific evidence

In light of the current omnipresence of neuroscience results in contemporary debates about psychological functions and the far-reaching conclusions that are often drawn from these by neuroscientists and lay-people alike, it may come as a surprise that Marr's influential methodology includes neuroscientific results only as the lowest of three – or even four[47] – different levels of analysis and explanation. As argued above, it was his disappointment with the relevance of neuroscientific evidence regarding

---

[46] Especially so-called enactive views of perception underline that human vision is often noticeable geared towards the specific actions or sensori-motor relations that a subject undertakes (cf. Noë 2004 ; Thompson 1995). Perhaps more plausible is a recent proposal suggesting that vision consists of a mutual interaction between early vision and top-down influences of object recognition in a Bayesian inferencing mode (Yuille and Kersten 2006).

[47] The four-level account presented in (Marr 1980) included a next to lowest level, consisting of assemblies of the basic components and activities that make up the lowest level. This concurs with the recursive decomposition of an explanatory mechanism that can be carried out according to the mechanistic explanatory methodology that will be discussed below in chapter I.5. An alternative addition of a fourth level is level 1.5 by Peacocke. This amounts to the insertion of a level between the computational and the algorithmic level of such theories that yield functional equivalence classes without yet specifying algorithms (Peacocke 1986).

feature-detectors that convinced him of the importance of multi-level explanations in cognitive neuroscience. For some time, a simple mapping of mental states to brain states was attractive to him, too: "[a]t the time the eventual success of a reductionist approach seemed likely." (Marr 1982 13). It was his realization that many neuroscientific results yield inadequate information about what contribution the neural activities make to the investigated function that disappointed him. Briefly stated, this led him to conclude that "reductionism does not imply constructionism" (Marr and Poggio 1977 2): it turned out to be impossible to derive from the computational theory of a cognitive function alone the information that would enable someone to build such a device, nor was it possible to assemble a functional system on the basis of evidence regarding potential neural components alone. For an adequate explanation of the perception of a visual scene, insights were necessary about the computations and the algorithms involved. However, this did not imply that insights about components were irrelevant. To begin with, in order to be performed, these obviously needed implementation in some form of hardware like an electronic device or a brain.

The lowest level of explanation concerns the 'hardware implementation' and has to answer: "[h]ow can the representation and algorithm be realized physically?" (Marr 1982 25, figure 1-4). This in itself is a complex issue, for to the same degree that algorithms differ between different animal species, different implementations are possible. To draw conclusions about particular implementations, investigations of the hardware of different species or other devices are required to provide information about: "how do transistors (or neurons) or diodes (or synapses) work?" (Marr 1980 199). The relevance of neuroscientific investigation for such information is unmistakable. However, at the same time some caution is in order.

The cash register discussed in section I.3.2 already suggested that addition and deduction can be performed in different ways, using different representation formats and algorithms. Such differences are not just important to consider for engineers designing a register, but also for neuroscientists investigating potential brain processes that can carry out those computations. Acknowledging this variability, Marr appears to subscribe to the assumption that cognitive tasks are *multiply realizable*, which implies that these tasks (or the 'psychological predicates' Putnam dealt with in his seminal paper (Putnam 1967)) not only allow description without recurrence to correlated brain-states but can also be correlated to multiple non-identical brain states.[48]

---

[48] As Marr realizes that frogs and rabbits may have different implementations of comparable computations, his position implies subscription to the idea of local multiple realizability of some functions as described in (Kim 1992). With respect to this Wimsatt remarks that it is: "entertaining to see philosophers of psychology act as if this characteristic is a special property of the mental realm" given its prevalence in nature (Wimsatt 2007 217).

At times it seems that Marr subscribes to a functionalist position, which attaches relatively little importance to the implementation level of the explanation of a particular cognitive function. This may be derived from statements like: "trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers. It just cannot be done" (Marr 1982 27). This sounds similar to Putnam's statement about the relative unimportance of implementation properties for the understanding of mental functions: *"What is our intellectual form?* is the question, not what the matter is. And whatever our substance might be, soulstuff, or matter or Swiss cheese, is not going to place any interesting first order restrictions on the answer to this question" (Putnam 1975 302, italics in original).[49] However, Marr does not deny the relevance of 'what the matter is'. On the contrary, for he continues his statement concerning the study of bird flight with the acknowledgement that here, too, the implementation level is relevant when other levels of analysis are attended to: "[i]n order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense" (Marr 1982 27).

In such a situation, knowledge about the implementation hardware can also provide us with clues about the nature of the algorithms that a device probably uses. For instance, since neurobiology shows that 'wires' or connections grow in brains rather quickly and cheaply and do so in three dimensions, parallel computing appears to be an attractive option, in contrast to the preference for serial computing by electronic devices and their two-dimensional printed circuit boards (Marr 1982 24). A confirmation of this preference for parallel computing that stems from such a fundamental principle of the biology of the brain has been the discovery of several streams of processing visual information in the human brain. This discovery is partly a result of neuroscientific studies and has had implications for the explanatory contributions that stem from other disciplines.[50] However, as was the case with the algorithmic level, where several options were found to be available for the same computation, the implementation level offers the investigator a similar room for choice. Not just because of the possibility that a particular brain or device may perform a certain computation in more than a single way, there is also a choice available in the granularity with which one looks at implementation. Focusing at a

---

[49] With the italicized question Putnam refers to the alleged Aristotelian preference for a formal explanation of the mind over a material one. He has in fact misjudged Aristotle's hylomorphism, which in this context led Aristotle to be a precursor of embodied cognition theory (van der Eijk 1997).

[50] It appeared that distinct kinds of information are subserved by different streams of visual information proessing (Goodale and Milner 1992), of which the number, their functions and their interaction remain a matter of debate (Creem 2001).

neural network level or at the finer-grained level of neurochemical interactions may give us different insights and accordingly different clues about the upper levels of the multi-level explanation of a function (Kosslyn and Maljkovic 1990).

As a result of the foregoing, we may conclude that explanations of cognitive functions involving the implementation level may be corrected without other aspects of the multi-level explanation automatically requiring correction as well.[51] Obviously, this should not lead to the conclusion that the implementation level does not really matter in explanations of cognitive functions. On the other hand, drawing conclusions on the basis of insights that focus on the implementation level alone is always rash and should be treated with extreme caution. To close this section on Marr's methodology, we will therefore focus on his proposal for relating these three levels as part of a comprehensive explanation.

### 3.4  A loose interdependency between levels

Marr's methodology is primarily based on the conviction that for many cases of understanding complex systems: "one must be prepared to contemplate different kinds of explanation at different levels of description that are linked, at least in principle, into a cohesive whole, even if linking the levels in complete detail is impractical." (Marr 1982 20). As noted above, the levels involved in this account do not refer to entities in an ontological sense or levels of a mechanism – to be discussed in chapter I.5 -, but rather to different 'perspectives' (McClamrock 1991).[52] The requirement to link such levels of description or understanding 'into a cohesive whole' leaves open the question how such linkage is to be carried out.[53]

Marr dismisses the option of always requiring unique connections between the analyses, since at each level of analysis of a given function "there is a wide choice available" (Marr 1982 25). In explaining a certain function at several levels of analysis, researchers must therefore leave room for the possibility that an alternative computation or algorithm is carried out during the performance of a function, which may require a different implementation than previously thought. The benefit of such multiple realizability prevalent at all three levels of analysis is: "that since the three

---

[51] Not surprisingly, however, especially Marr's ideas about object recognition have been criticized with respect to the computation and algorithmic levels and with respect to its implementation (Glennerster 2007).

[52] MacClamrock convincingly argues that Marr's levels have no ontological import but should be considered as three different perspectives on any given function or component function (McClamrock 1991).

[53] The following rendering of Marr's account leaves out this question of integration of the levels: "levels of analysis concern the conceptual division of a phenomenon in terms of different classes of questions that can be asked about it" (Churchland and Sejnowski 1988 741). The levels in Marr's approach are more than just a method for framing questions or a heuristic, as will be argued.

levels are only rather loosely related, some phenomena may be explained at only one or two of them" (Marr 1982 25). After all, the properties or constraints that appear at a particular level do not all transpire to the other levels of the function. For instance, although a computer may use completely different algorithms from a human brain, both systems may still perform the same task (Marr 1982 27). If we want to compare different systems, it is therefore very important to target the relevant level of analysis for such a comparison.

The loose interdependency also has as an advantage that the methods and results pertinent to all three levels do not always have to progress simultaneously. Equally, it makes each level relatively immune to flaws that are present at other levels, just like Marr's dissatisfaction with the neurophysiological strategy of searching for feature detectors inspired him to put more efforts in developing an adequate computational or algorithm theory for the understanding of such neurophysiological findings. Understanding a particular algorithm can at times help to make progress in uncovering the associated mechanism, or vice versa. Their 'loose independence' does not exclude the presence of mutual constraints, where results at one level can guide research at another level, for example: "the form of a specific algorithm can impose strong constraints on the mechanisms, and conversely" (Marr and Poggio 1977 12) – the discussion in the previous section of biological wires and a preference for parallel computing confirms this observation.

Nonetheless, the role of mutual constraints between levels of explanation is less prominent than it is in the mechanistic explanation of a cognitive function. As we will see below, this is due to a different take on the idea of 'levels' itself. For Marr, the levels of explanation or analysis are perspectives on a particular function or component task of that function. Even though he emphasized the relative independence of these perspectives, it is important to note that one of the major contributions of his work has been to demonstrate the feasibility and necessity of integrating different perspectives: "Before Marr, researchers who studied artificial intelligence were concerned with 'disembodied' information processing, and paid scant, if any, attention to the brain; and researchers who studied the brain were concerned with neuroanatomy and neurophysiology, and paid little attention to formal analyses of what the circuitry does. These communities are no longer isolated from each other, and this ultimately may be Marr's greatest achievement" (Kosslyn and Maljkovic 1990 250).

A similar integration of perspectives likely benefits the investigation and explanation of other cognitive functions, like singing or action determination. In those contexts as well, we may hope for a loose independency yet with some interdependency between the different levels of analysis – or disciplinary perspectives

- involved. To make our hopes more robust, however, it is useful to end this chapter with a short evaluation of Marr's approach.

## 3.5  Modularity and some limitations of Marr's methodology

Even though Marr strongly defended his approach, he was quite aware of some of its limitations. He avoided taking an extreme position, for example, with respect to the predominance of one level of analysis over the others. Having distinguished different levels of analysis and associated these with different strands of research, he demonstrated how a careful navigation between these levels bolsters our understanding, whereas partisanship for a particular level risks unnecessarily weakening it.[54] Instead, the combination of these levels is what furthers comprehensive understanding of a particular function or function component. What such a combination seems to require, however, is the relative isolation of such a function or function component. Only on the basis of a certain 'modularity' of a given object of investigation can researchers apply their multi-level analysis to it, Marr assumed. This he thought to be a prime reason for the limitations of his approach. We will elaborate on this assumption of modularity and at the end of this section link it again to his advocacy of loose independence between levels of analysis.

As noted earlier, Marr's main research was on early vision, a stage of visual information processing which he believed to be relatively insensitive to other functions of the organism or to later stages of visual perception.[55] This assumption met with criticism on both theoretical and evidential grounds. Marr was not naïve in this matter and realized that the assumption might turn out to be inappropriate. Supporting the assumption, though, was the consideration that the evolution of a complex function would generally benefit from avoiding the situation in which it was vulnerable to many small and local changes in the overall system.[56] Correspondingly, he assumed that the evolutionary development of such a function could better not interfere with earlier developed component functions.[57] Obviously, later evolved

[54] Marr's emphasis on the computational theory has nonetheless made some authors argue that he denied the implementation level's relevance (cf. Glennerster 2002 ; Polger 2004 ; Sun, Coward et al. 2005). We think that (Kosslyn and Maljkovic 1990) are more correct in stating that Marr's works demonstrate an imbalance that he theoretically did not justify.

[55] This assumption of early vision's peculiar nature was disproved many years ago. See e.g. the more recent account of V1's function and place in the architecture of vision in (Ng, Bharath et al. 2007) which includes not just feed-forward but also several feed-back relations to V1.

[56] This estimation of interaction may be due to Marr's idea that the stages make up a sequence-like process, where information must always pass through these stages (Glennerster 2002).

[57] The correspondence of evolution with modularity is indeed plausible, as is argued in (Wimsatt and Schank 2004), although they emphasize the importance of distinguishing between different forms of modularity in this context.

functions could take up tasks in addition to tasks that had already been performed by functions that evolved earlier. Indeed, he believed that: "as evolution progressed, new modules came into existence that could cope with yet more aspects of the data" (Marr 1977b 41).[58] To capture this idea of developing functions more generally, he formulated a 'principle of modular design' (Marr 1976 485; Marr 1982 102) which states that breaking up a large computation into small and independently executed ones makes it more robust.[59] In addition to its robustness, such a structure would also enhance the speed and fluency of the process,[60] which could be hampered by hierarchical interactions (Marr 1976 502).[61] In the second part of this dissertation, we will discuss empirical hypotheses regarding the prevalence of modular structure in a cognitive system like the brain as a result of its evolution and ontogenetic development. For it has been argued from various perspectives – computational, evolutionary, developmental, to name a few – that such a structure yields several benefits to the organism or system that is equipped with such a structure.

However, apart from such benefits, Marr also had some epistemological considerations, believing that a truly complex system with many interactions between different stages of its processing trajectory might be impossible to understand.[62] Indeed, in the 'Conversation' added to his book on vision, he even admits that his approach would fail with: "Systems that are not modular. Things like the process by which a chain of amino acids folds to form a protein - that is to say complex, interactive systems with many influences that cannot be neglected."

---

[58] As we will discuss more below, the 'Massive Redeployment Hypothesis' argues that evolution plausibly also proceeds by re-using components that are already in place: evolution proceeds by redeploying and adjusting older components of an organism, suggesting integration rather than mere modularity as the modus operandi of many components (Anderson 2007).

[59] Marr did not articulate the notion of modularity as Fodor did around that time. Still, there are clear affinities between both notions. Indeed, several of Fodor's defining properties of modularity would apply to Marr's theory of early vision. The Fodorian characteristics of a modular system that seem to apply to early vision are its being a mandatory input system, its providing 'shallow' outputs to later stages of vision, and its being a system that is hardly accessible by other cognitive functions and working fast. Other Fodorian properties of modularity, like the informational capsulation of the coupled visual-motor system or the association with a fixed neural architecture and with specific failures do not have direct parallels in Marr's work or are not discussed there (Fodor 1983).

[60] When discussing language, Marr expected it to be modular in structure because of the characteristics of its 'fluency', its smooth continuation and the absence of consious attendance or interference (Marr 1982 356). However, some hesitation is aired in that context: "It's not clear, and some claim it's inherently not modular and should be viewed much more heterarchically" (Marr 1982 356).

[61] As we know, quite the contrary –that sensori-motor coupling makes vision an easier task- has been extensively argued in i.a. (Clark 1997 ; Noë 2004 ; Thompson 1995). Marr may have underestimated the benefits that a hierarchical structure yields to a complex and dynamic system. Or, to be more specific, the benefits that a heterarchical structure yield – see footnote 96 on heterarchy.

[62] Simon confesses that he is unsure about the chicken or the egg regarding his epistemological preference for a particular structure of a complex system and his argument that such a structure must have been more profitable for the system during its evolution (Simon 1962).

(Marr 1982 356).[63] Similarly, he adds, syntax appears to be almost modular although it may have sparse interaction with semantics, necessary to narrow down syntactical information in some cases. From that short discussion we can derive that Marr fears that a system or function which is not part of a strictly hierarchically structured process may be very hard to analyse and explain. Such a function, whose processes would then be partly determined by continuous interactions with other functions, is hard to study and then explain in isolation from those other functions. Due to its complex structure, a function could turn out to be modulated by other functions or component functions at stages that were originally assumed to be irrelevant to it. We may be forced to expand the function under study and include ever more functions in it that were previously thought to be distinct from it – making it ever harder to analyze and explain it.

The result of these considerations regarding a function's consisting of isolable modules corresponds in a certain sense to the earlier considerations about the relations between the levels of analysis in Marr's approach. As much as the computational, algorithmic and implementation levels of analysis of a function were found to potentially constrain each other's results and guide each other's investigations, they would still allow researchers to offer separate explanations. Only because of this separation of different levels of analysis or different perspectives on the function under study, would flaws or misunderstanding at one level not automatically affect the results at other levels. We may conclude that in Marr's view, the robustness of a system or function with a modular structure corresponds with the robustness of the scientific results of its study when these results do not rely too much on each other.[64] Given the situation in cognitive neuroscience nowadays, in which functions are allegedly subserved by heterarchically structured neural networks and in which notions like embodied or enactive cognition reign, one may well wonder why Marr's approach still receives the attention and support it does. Indeed, when Ochsner &

---

[63] Elsewhere Marr discussed the distinction between two types of theories that are employed for solving problems in artificial intelligence. Type-2 solutions describe complex and interactive problem solving processes. At the time (in 1977), Marr advised not to concentrate on such difficult problems, while noting that many authors focused on problems that humans perform poorly even though we may readily understand these problems intellectually. For instance, they focused on arithmetic, even though Marr does not believe that we mentally perform along the lines that arithmetic problem solving is described. Ironically, he concludes that "one is left in the end with unlikely looking mechanisms whose only recommendation is that they cannot do something we cannot do" (Marr 1977b 45).

[64] It has been argued that multi-level systems can also be analyzed and explained in terms of a hierarchical model structure, where a model that accounts for a sub-system at one level can constrain the number of plausible models at another level. The authors explicitly parallel their hierarchical model structure with Marr's multi-level approach (Meeter, Jehee et al. 2007), although in our view the comparison denies the quite distinct nature of Marr's implementation level from the other two levels.

Kosslyn list their 'five general points about cognitive functions' as they are studied in cognitive neuroscience, it appears as if their list contains several objections to Marr's approach. This holds particularly for those referring to the highly interactive nature of the neural networks that make up cognitive functions, which are implemented in distinct brain areas and are processing information not just serially but also in parallel (Ochsner and Kosslyn 1999 354). Similarly, it has been argued that both the phenomenology of perception and the apparent dynamical processes that underlie perception (Borrett, Kelly et al. 2000) are at odds with the modularity assumption, which we found to be prominent in Marr's approach. The development of parallel distributed processing and the discovery of such processes in the brain's networks have also suggested that some of the advantages that Marr attributed to modularized systems – like speed, robustness, and evolvability – can actually be exhibited by those neural networks.[65] However, the suggested opposition between the assumption of modularity and the parallel distributed nature of neural networks itself deserves to be considered along the lines of the approach Marr defended.

According to that approach, we should not be surprised to find that a computational theory of a given function – and its psychophysical or cognitive psychological investigation – involves a certain degree of modularity of that computation or the algorithm that performs the computation. Nonetheless, the implementation theory of the relevant computation or algorithm must be assumed to be only loosely related to theories regarding those other levels, leaving ample room for implementation by a network structure that shows a more limited form of modularity. Assuming otherwise, that an implementation by a parallel network would contradict a modular structure of a function, is usually based upon the unwarranted conflation of the anatomical structure subserving a function with its functional structure as described by its computational theory.

Even though the assumption that cognitive functions are subserved by the parallel distributed networks that are prominent in the brain has gained ever more prominence, the distinction between functions and components functions has survived. More and more, the results of evolution, development and learning are being described in terms of the acquisition of specific modules or the development of a function's modularity. In Part II we will see how and why this is the case, when we consider

---

[65] Paul Churchland has argued extensively about the characteristics of such parallel distributed networks or connectionist approaches for a variety of cognitive functions in his (Churchland 1995), while Patricia Churchland many years earlier paved that way in her seminal (Churchland 1986) in which she notes that the parallel distributed networks approach grew partly because of the unsatisfying results with serial computational approaches. Marr's approach was more of the latter fashion, indeed. Had Marr lived longer, he probably would have agreed with that.

why it is plausible for dynamic and evolving systems to develop intermediate stable forms (Simon 1962), generative entrenchments (Wimsatt 1986), or kludges (Clark 1987). We will explore empirical cognitive scientific theories about 'modularization' of cognitive functions that takes place during child development (Karmiloff-Smith 1992) and the formation of distinct processes or functional systems – 'kludges' - for extensively practised cognitive functions.

In sum, even though several of Marr's specific theories concerning – components of - vision and some of his methodological assumptions do not find adherents anymore, some of the general principles of his approach are still considered relevant and beneficial to a 'flourishing' neuroscience (Rolls 2011). As we turn to the next methodologies to be discussed in this part, the reader may discern how Marr's approach has been integrated in other and more recent ones. Moreover, since these approaches are being applied to cognitive functions that are considered to be 'higher' and more complex than the early and middle stages of visual information processing to which Marr limited his research, their discussion prepares us for the investigation of the domain of this dissertation, concerning the determination of human action.

## 4  MODEST, ALL TOO MODEST: THE SEARCH FOR NEURAL CORRELATES

Singing, we have seen earlier in chapter I.2, is impossible to define in such a way that it delineates a conceptual space in a comprehensive, consistent and coherent sense without remaining ambiguous cases or blurred boundaries between, for example, vocal signalling, infant crying and novice vocalizations. In the previous sections a somewhat more liberal approach aiming to combine several perspectives on a given function was discussed. In agreement with the earlier approach that depended upon conceptual analysis, Marr's approach emphasized the relevance of an account of the task being studied, while leaving room, however, for there being several such accounts instead of only one. Constraints on the task definition or its computational theory could subsequently be derived from the algorithmic theory, devoted to answers concerning the kind of tasks involved in singing, its particular goals, the distinction of subtasks that merit separate study, and the like. Similar relations between levels of analysis may involve the third level of analysis, focusing on the neural implementation of the function. The plausibility of a particular answer can be partly based upon its coherence with the results of investigations of the task at the other levels or other perspectives involved.

As singing is a complex and difficult task, it is convenient to use a strategy of divide and conquer. Through dividing it into several subtasks that can be investigated separately, researchers have made the problem more manageable. Indeed, as we have seen in section I.3.5, a working hypothesis behind this strategy is that tasks and subtasks are to a large extent modular in nature. Basic to this approach is our distinction of the functions involved and their subtasks, and our subsequent finding of empirical support for that distinction. Importantly, we should not overlook the possibility that our classificatory distinction leads us astray and eventually turns out to be implausible.[66] On the other hand, not making any classificatory distinctions at all may result in a situation where we have no difficulty in expanding our research topic 'signalling-crying-singing' by including ever more social communicative functions in it, like gesturing, sobbing, shivering, and so on. Expanding the computational theory of our research topic would make it ever harder to distinguish a specific neural implementation for this function other than the brain as a whole. Designing a single

---

[66] That classifications and taxonomies can misguide us in dividing a particular set in two or conversely lumping together two sets that had better remain apart has been a topic of discussion since Plato and Aristotle put emphasis on definition as a scientific principle. A lesson to be drawn from this is not the entire rejection of definition – which would leave us with even greater problems – but to allow for some pluralism corresponding with different questions, as is convincingly argued in (Dupré 2001).

algorithm that would allow the brain to perform this comprehensive function in all its distinct expressions would be even more difficult. In sum, some classification is necessary as a starter for research, as is some kind of modularity – even if they need to be corrected at a later point.

The tension between the investigation of a complex and comprehensive function that is hard to study as such, and the assumption of it being subserved by subtasks that do allow separate study is prominent in consciousness research – which is arguably a more prominent, larger and even less defined domain than the study of singing. For the next methodological approach, we will turn to the study of consciousness and in particular to the study of so-called Neural Correlates of Consciousness (NCC). This approach does not present a method to study the comprehensive function directly and instead suggests studying only particular components or particular phenomena related to consciousness. It allows researchers to focus on a particular phenomenon that is related to consciousness or that represents a specific form of consciousness, while remaining relatively silent about consciousness as a whole. When some form of decomposability is assumed, the distinguished components are not separated as modules from the comprehensive function, as was the case when Marr studied stereopsis as part of early vision. What is important is whether the levels of analysis that Marr distinguished play a role in this method of the study of NCC. For example, is there a requirement of explicitly articulating computations or algorithms, such that it enables researchers to relate different NCC's to each other? Or is the focus different, perhaps on the implementation level? Is there an interest in constraints that mutually help to determine the theories at different levels of analysis, as Marr suggested?

## 4.1  Identifying a minimal yet sufficient neural correlate – of what?

Considering the great variety of phenomena and concepts associated with the 'last surviving mystery' of consciousness, Dennett raises the possibility that: "[p]erhaps the various phenomena that conspire to create the sense of a single mysterious phenomenon have no more ultimate or essential unity than the various phenomena that contribute to the sense that love is a simple thing" (Dennett 1993 23).[67] If that is the case, then it seems that the preliminary requirement of defining a cognitive task or providing a task analysis is to be avoided. Instead of seeking to identify and explain such a simple thing, NCC researchers have been focusing on a series of particular

---

[67] Classifications of consciousness are rich and to some extent divergent among authors – compare those offered in the invited review of (Zeman 2001), according the conceptual analysis of (Bennett and Hacker 2003), or in the argument for investigating computational correlates alongside neural correlates of consciousness in (Atkinson, Thomas et al. 2000).

phenomena generally considered to belong to the class of events to which the concept of consciousness applies. To this end, a method has been developed which allows them to study a series of cognitive events in an animal or human subject such that they assume to be capable of distinguishing these events according to their property of being conscious, or not. Let us explain this curious method.

As we have seen, explanation can occur in various ways. Not only can we facilitate the explanation of a phenomenon by taking up different perspectives or levels of analysis, or by dividing the phenomenon into components that facilitate the project of investigation and explanation, as we have witnessed above. Another method that can simplify explanatory work is to describe a phenomenon not by providing a comprehensive account of it, but by contrasting it with a comparable phenomenon which differs from it in a specific respect: why does X occur instead of Y (Ruben 1992)?[68] Instead of explaining the riding of a car, we can explain the contrast with the car's riding back and forth or we can aim to explain the difference between singing mere melodic lines and singing a song.[69] When offering such a 'contrastive explanation', it is much easier to decide which information is relevant and which is not, since we aim to explain a specific difference between comparable phenomena and not a single phenomenon in its entirety.[70] For example, contrastive explaining of the car's riding direction will direct us to the gearbox while taking the activities of the motor, wheels and breaks for granted. Explaining the difference between vocalizing and singing may inform us specifically about the neural activities that correlate with semantic processes in vocal expression.

Returning to consciousness, there is broad consensus that consciousness refers to a phenomenon or set of phenomena that is absent when subjects are asleep, unconscious or in coma and which reappears when they wake up.[71] However, that transition is too intricate and ramified to allow detailed study, as it encompasses several distinct sensory and cognitive functional differences. Accordingly,

---

[68] Ruben warns against overstretching the importance of contrastive explanation: he argues that not all explananda are contrastive and that contrastive explanations are not all equally apt for traditional non-contrastive explanation (Ruben 1992 39-44). What is relevant in the present context, though, is that contrastive explanation corresponds particularly well with the subtraction method in many neuroscientific experiments, in which the neural activitions correlated to two - test and control - conditions are compared.

[69] Analogously, the causal relation to be investigated is not just a relation between a cause and a particular phenomenon, but between two comparable causes, where one of them is causing that particular phenomenon while the other is responsible for another phenomenon (Schaffer 2005).

[70] Indeed is such contrastive explanation quite common in consciousness research, probably because of its relative modesty as (Hohwy and Frith 2004) argue.

[71] Revonsuo has a different approach, holding that the dreaming brain is similar to consciousness with respect to phenomenal awareness and thus a good place to start consciousness research research (Revonsuo 2000).

researchers have been keen to identify cases in which a similar, specific and more minute transition is observable. That transition should consist of at least a single component associated with consciousness, that is to say with conscious experience. This component requires subjects to report their experience behaviorally or verbally, excluding sleeping or comatose subjects who are principally unable to report on the transition. Indeed: "the fundamental methodological problem faced by any rigorous research program on consciousness is the subjectivity of the target phenomenon" (Metzinger 2000 1). Given this and in view of Dennett's reminder that we should not focus our study of consciousness on a single phenomenon, the question is how we can isolate specific and minute transitions from non-conscious to conscious experience that are reportable by subjects. If researchers gather evidence about a number of such transitions, hope is that together they may teach us something about the structure and neural implementation of consciousness more generally.

Probably the first study to focus on such a minute yet reportable transition was with monkeys, investigating the 'Neuronal Correlates of Subjective Visual Perception' (Logothetis and Schall 1989). Interestingly, it did focus on visual motion yet not merely on the task of visual motion processing according to Marr's approach. Instead, it targeted the transition of visual motion perception into and out of reportable experience. It consisted of a binocular rivalry task, in which the eyes of the monkey were continuously presented with two different stimuli, which could not be perceived simultaneously. If the left eye is presented a downward movement and the right eye an upward movement, the reportable perception will alternate between the two stimuli, even though they themselves remain stable. The question of the authors was whether the neurons involved in processing visual motion are the same that correlate with the reportable perception of movement. They found that indeed there were single cells whose firing rates correlated with the subjective visual perception as reported by the monkeys.

This correlation between recorded firing of specific neuronal cells and the subjective visual perception is fascinating but still leaves room for alternative explanations, as we are not sure whether transient perception correlates with other neural activities as well.[72] Echoing Marr's worry that his approach of investigating components of vision could fail if vision turned out to be a highly interactive process, the authors note that the correlation itself needs further interpretation because: "the

---

[72] (Noë and Thompson 2004) argue that this analysis of the binocular rivalry task obscures the fact that the experience of this rivalry – or similarly in perception of a bi-stable figure like the Necker-cube – is a perceptual experience in itself, which is temporally extended and encompasses the two distinct percepts between which that rivalry exists. Although we concur with their observation, it does not exclude the option to focus on the rivalry as such.

perception-related modulation observed in these neurons may be a result of feedback from higher centers" (Logothetis and Schall 1989 763).[73] It would require additional research to exclude the possibility that this reportability may be due to such a higher center and not rely just on the neural areas investigated in this study. In its aim to investigate a minute and specific transition of a percept in and out of a reportable status, this approach seeks correlating neural activities that are as specific for that transition as possible.

The method of searching for a neural – or neuronal, as it is sometimes called – correlate of consciousness has become ever more articulated since this first monkey study. The general hypothesis behind such research has remained that: "it is useful to think of consciousness as being correlated with a special type of activity or perhaps a subset of neurons in the cortical system," without assuming that consciousness is a single phenomenon, nor that its neural correlates always have to be identical (Crick and Koch 1990 266). In fact, researchers are looking for a specific neural correlate that is not a mere side-effect of a case of conscious experience but that is itself responsible for a specific case. An often-used definition captures the NCC research goal more technically: "An NCC is a minimal neural system N such that there is a mapping from states of N to states of consciousness, where a given state of N is sufficient, under conditions C, for the corresponding state of consciousness" (Chalmers 2000 31). In the definition, the clauses of minimality, sufficiency and mapping relation are obviously the most remarkable. However, given the fact that minimality and sufficiency conditions are not exceptional for this research method, we will not discuss them further.[74] The other condition, that one is looking for a mapping relation, is relevant here and deserves further scrutiny.

The mapping relation may remind us of the 'loose relations between levels' that Marr defended (Marr 1982). In the present context, the result of the search for an NCC is not the development of a causal model but a result that allows researchers: "a mapping from states of N [ = a minimal neural state, MK] to states of consciousness" (Chalmers 2000 31). Such a mapping or correlation between neural and conscious states gains in relevance as the specificity of those states becomes more and more

---

[73] As discussed more below, recurrent processing has more recently indeed been found to be correlated with reportability of visual stimuli – see e.g. (Lamme and Roelfsema 2000).

[74] Obviously, considering the brain as a whole as an NCC is uninformative, which is why: "it makes sense to 'start small' in the search for an NCC" (Chalmers 2000 32). However, if a particular NCC were not sufficient for producing a specific conscious experience, the explanation would probably still be incomplete, as other neural activities would be equally necessary. Single cell studies like the one by Logothetis e.a. mentioned above exemplify this strategy (Logothetis and Schall 1989). fMRI activations patterns are less specific but may in turn inform us about task modulations by less relevant processes related to motivation, attention, learning and memory, as one of the authors argues (Logothetis 2008). Excluding such 'neuromodulations' from the NCC is an important methodological task.

explicit. One way to achieve this is to focus not on consciousness generally, but on a specific content of consciousness, as in the study of binocular rivalry, when the aim was to identify the neural correlates of the reported visual percept. However, this requires a preliminary basis for identifying a particular state of consciousness as being indeed just that: a state of consciousness.

Such an identification is not as simple as it seems. In the case of binocular rivalry, for example, one could ask what exactly the unit of consciousness or the precise state of consciousness is for which researchers have been trying to identify a correlating neural state. Is the conscious experience of binocular rivalry in fact a couple of experiences that alternate between two different conscious states, or is it the singular and comprehensive conscious state of experiencing the switch between percepts? Should we identify the neural correlate of two different, snap-shot like conscious states – involving two non-identical and non-overlapping percepts – or should we look for the NCC of a temporally extended conscious state which includes the alternation between two different percepts (Noë and Thompson 2004)? Arguing in favor of the latter, Noë & Thompson emphasize that researchers have to pick the right level of analysis when looking for an NCC: "the content of perceptual experience is personal-level content, not subpersonal-level content" (Noë and Thompson 2004 18).

The distinction between a personal-level and a subpersonal-level of analysis here distinguishes between a phenomenological account and a neurophysiological account of an experience. The methodological distinction between levels proposed by Marr may be more informative in this case: what is the computational theory or task analysis of the conscious state for which researchers are trying to identify an algorithm and implementation theory?[75] Is it plausible to expand the computational theory of bistable percepts such that it includes both alternating percepts and their alternation, or disregard the bistability as such? Obviously, both cases will yield different results in terms of the correlated implementation theories, especially in terms of the temporal structure of the alternation. When bistability is disregarded, the temporal dynamics of alternation may be left out of the equation completely – simply as a consequence of a particular task analysis of bistable perception.[76]

---

[75] The authors reproach the NCC methodology for not providing us with a causal explanation of an experience by a neural state (Logothetis 2008). However, a reliable correlation between two states at different levels of analysis can in itself be considered a part of a scientific explanation. Moreover, it is not so much a causal but a constitutive relation between the neural state and the conscious state that we should expect in this case. This is acknowledged in the mechanistic account of explanation, discussed in the next section of this Part I.

[76] A framework that aims to explain the fact that it is continuity and discreteness together that make up our phenomenal experience is offered in (Fingelkurts and Fingelkurts 2006). These authors also contend that the phenomenal level at which continuity and discreteness are both experienced should be the starting point for the development of an explanatory framework.

The foregoing may warn us against avoiding the issue of first analyzing the function or task to be investigated. This warning was already expressed in Marr's methodology and therefore not new to us. Nonetheless, the NCC approach to some extent avoids the development of a computational theory, since consciousness is accepted as being a research topic so hard to determine that the hope is that its delineation may be reached by evidence converging on one or more neural correlates of states accepted as being conscious – even without a more technical analysis or definition of those states. This hope has led to another strategic choice, aimed at limiting the phenomenon for which an explanation is required. It has been generally accepted in NCC research to distinguish between research of the 'background state of consciousness' like wakefulness or dreaming and research of specific contents of consciousness (Chalmers 2000). Even though the two are likely related, their investigation is facilitated by separating them. Clearly, the experience of binocular rivalry or bistable percepts is possible only on the condition of being conscious and awake, but these aspects of the background state are left out of the explanation of the contents of the experiences. Although the reason for the omission may be convincing, the consequences should not be neglected.

## 4.2 Further limiting the phenomenon and its correlates

We started the previous section by quoting Dennett's consideration that the phenomenon of consciousness is so complex that we may be well-advised to accept that it in fact consists of various phenomena that do not conceal a single one (Dennett 1993). The search for a NCC concurs with that consideration, as it usually aims to determine a minimal neural correlate for a particular conscious content. Given the wide-ranging variety of phenomena associated with consciousness, this strategy avoids both the conceptual and the empirical challenge of presenting a unifying definition of some sort. Rejecting the 'simple thing' assumption about consciousness also provides a practical advantage, since it allows researchers with as many possible ways of studying and accessing consciousness as there are associated phenomena (Churchland 2005).[77] Even though these problems appear particularly intricate for the phenomenon of consciousness, several of the lessons learnt from these regarding the methodology of research are relevant in other domains of cognitive neuroscience as well.

Common to these lessons is that limiting the phenomenon under scrutiny increases our potential for explaining it. This obviously comes at a cost, however. In the previous

---

[77] Obviously, recognition of the variability and the multifaceted nature of consciousness does not preclude researchers to attempt a unified theory of consciousness. However, as Seth and others argue in presenting their framework, such a theory should include measures for both various quantitative and the qualitative aspects of consciousness (Seth, Izhikevich et al. 2006).

section, contrastive explanation was invoked, as it allows a focus on explaining the difference in reportably perceiving either picture of a binocular rivalry task, without having to explain all neural conditions necessary for perceiving any picture at all. In the domain of consciousness studies, a similar distinction has been made: "between the neural correlate of background state of consciousness (wakefulness, dreaming, etc.) and the neural correlate of specific contents" (Chalmers 2000 33). Of course, it should be realized that the NCC resulting from the study of a specific content of consciousness may produce that state only against such a background state, while the conditions of this background state are not themselves included in that particular NCC. Accordingly, the NCC was defined as a minimal and sufficient neural system for producing the contents of a very specific conscious state – minimal, that is, against the presence of such a background state. Since this background state remains unspecified, the interpretation of the associated NCC is troublesome, because such an NCC by itself is not really sufficient for any conscious experience to emerge. Indeed, "momentary activations in local neural populations in visual system, which the definition of the NCC is carefully constructed to pick out, are not sufficient for any actual experience had by any actual subject"(Neisser 2012 683-684).[78] In spite of these reservations, it is not uncommon in cognitive neuroscience research to limit the phenomenon under study and its explanation in this manner.

Not surprisingly, the strategy has raised serious criticism. Following their line of critique aimed at the delineation of binocular rivalry, the same authors draw attention to the background state and the activities of an animal or human subject that has conscious perception: "if perceptual content depends on the skillful activity of the whole animal or person, making use of its capacities for eye, head, and whole body movements, and for directed attention, then it becomes questionable whether there is any such thing as a minimal neural substrate sufficient to produce experience" (Noë and Thompson 2004 17).[79] If this should be taken to suggest that an NCC for a particular experience should include all correlates responsible for this rich texture, then it is unhelpful since that would likely involve the most part of the brain. Again,

---

[78] Consider for example Searle's conception, involving a rather different task analysis – or computational theory, in Marr's words – of consciousness: "[w]e should not think of perception as creating consciousness, but as modifying the preexisting conscious field" (Searle 2004 81). In addition, he writes that the NCC thesis regarding correspondence between states is in fact a trivial one.

[79] As Noë and Thompson remark –while referring to Varela- this line of research is impossible to carry out "independent of the sensorimotor context of the animal as a whole" (Noë and Thompson 2004 13). That brings them to the conclusion concerning the NCC approach that "[w]hat makes this account internalist is that it views the experience as supervenient on neural processes alone" (Noë and Thompson 2004 20) and consequently independent of interactions with the body and the external world. However, as imaging techniques are used to investigate whether apparently unconscious or comatose patients are in fact capable of some kind of voluntary responses, such bodily interactions are circumvented indeed (Schnakers, Perrin et al. 2008). Unsurprisingly, this approach of sidestepping normal behavioral evidence of consciousness by using imaging evidence has been criticized by Hacker et al. who fears that at some point a neural correlate may even be accepted to overrule behavioral evidence (Nachev and Hacker 2010).

however, we can just accept that skillful activity and directed attention are generally involved and still aim for the limited NCC that would allow us to merely distinguish between the specific contents of two different conscious states. Obviously, this additional limitation does not make the interpretation of the function of the NCC any easier. A third limitation deserves mention.

As much as NCC research does generally depend upon disregarding the background state of consciousness and avoids a definition or task analysis of conscious experience, it must also 'screen off' the resulting NCC such that it captures the relevant condition. For if we find some activation – via single cell recording or fMRI, for example – in correlation with a particular consciously experienced content, we still must decide which feature of that activation is responsible for the correlated conscious state. We cannot just be satisfied by pointing out the involvement of a neural area or network, for it may still not inform us what property of that area yields the conscious state. As a case in point, Hardcastle observes that there is an ongoing dispute between 'smallists' and 'largists' about how to define the neural correlate. On the one hand, there are those who expect that small scale quantum effects are responsible for consciousness, on the other hand there is increasing interest in large-scale dynamics within the brain (Hardcastle 2000).[80] So is it the particular small scale quantum effects in specific neural cells that are relevant, or are these common to all cells in all conditions and thus irrelevant, whereas it is their firing in synchronicity with other cells that determines the conscious state? Researchers hope that this uncertainty can be settled by comparing several different NCCs, assuming that this will eventually tell us which of their properties overlap in most cases, and which do not.[81]

In sum, the search for NCCs depends upon strategic limitations in several senses or directions: limiting both the target phenomenon of consciousness, distinguishing it from a background state of consciousness and from other relevant activities and cognitive functions of the subject. A way to do this can be to engage in contrastive explanation of why a particular conscious state obtains instead of another. In another direction, the limitation involves the determination of the relevant neural state that correlates with the conscious state under scrutiny. These limitations can be found in

---

[80] Hardcastle mentions the necessity of a 'screening-off analysis' which amounts to determining the probability of potential causes A and B co-occurring with the event, and comparing these with the probability of the event itself, in order to decide which cause is doing the actual work (Hardcastle 2000). As far as we can see, this works best with a rather simple branching structure of causal relations, but is less adequate in a complex and interacting system including recurrent processes.

[81] For example, reviewing existing evidence for NCCs of conscious visual percepts, Lamme argues that it is a specific - recurrent – type of processing that is the relevant and overlapping NCC (Lamme 2006). More recently, is has been suggested to look for recurrent processing in vegetative, anaesthetized and dreaming subjects as a way of looking for residual conscious experience (van Gaal and Lamme 2012).

other strategies of research as well, to be sure. However, they are relatively prominent in NCC research, which is the reason why they were discussed here. Moreover, they enable us to formulate more specifically what in fact we expect from an explanation and in what sense a NCC may not be adequate.

## 4.3 Lessons on explanation drawn from the NCC research

Despite considerable research efforts, consciousness remains an elusive phenomenon. This is due to several factors. As mentioned earlier, researchers have not been able to agree upon a particular definition of consciousness. In fact, it has been argued that consciousness should not be considered as sharply distinguishable from non-conscious processing, but as a gradual phenomenon instead (Cleeremans and Jiménez 2002). Similarly, a different taxonomy has been proposed in order to explain and distinguish different forms of conscious and related phenomena (Dehaene, Changeux et al. 2006). Divergence has not only affected the definitions of consciousness or the distinction between conscious and other states, but there is also no consensus about the criteria for empirically determining whether a specific instance represents consciousness (cf. Lamme 2006 497, fig. 1 on measuring conscious visual experience). Not surprisingly, consensus regarding the functional role of consciousness is equally lacking, as much as it is absent with regard to neuroscientific explanatory accounts of it. So how do researchers respond to this situation and what progress in explaining consciousness do they aim at?

First, researchers generally accept the diversity and heterogeneity of conscious phenomena and experimental approaches to consciousness, as we mentioned in the previous section. Doing so, they fail to comply with the requirements of preliminary conceptual analysis or task analysis, as we saw above. Indeed, NCC researchers don't assume their object of research to be a 'simple thing' and appear to be relatively liberal in accepting phenomena into the domain of consciousness research, expecting that with such a wide net they will eventually be able to distinguish relevant from irrelevant phenomena as suggested in (Churchland 2005). And of course, they subsequently develop strategies to exclude phenomena from that net while simultaneously trying to find some overlapping property of those phenomena that are accepted as belonging to the space of consciousness.

Since evidence of consciousness in everyday life or in research generally depends upon behavioral or verbal activities, in many cases consciousness has become associated with properties of behavior, reflection, language, and so on. The second strategy contributing to progress in research therefore consists of showing that consciousness does not always depend upon the co-occurrence of these phenomena

and can even be dissociated from them. For example, a recent review of NCC research concludes from evidence that conscious experience "does not require sensorimotor loops involving the body and the world, does not require self-reflection (or language), and does not reduce to attention" (Tononi and Koch 2008 240). Such a dissociation of these types of response from consciousness itself is a useful result. It should not be forgotten, however, that any such dissociation is only possible on the basis of some sort of a previously accepted delineation of phenomena that attest to consciousness. Then again, the first strategy was to be liberal in accepting phenomena in the domain of consciousness research. So how is a limitation eventually brought about?

Apart from being liberal in admitting phenomena to the domain of consciousness research and then trying to dissociate these phenomena from other, associated cognitive functions, a third strategy is to look for some overlapping property on which the remaining phenomena converge. It is not surprising that such a property will be relatively abstract or general. For example, one proposal that finds widespread recognition is that consciousness is related to information processing.[82] On top of this - not unexpectedly in computational neuroscience - several authors who employ the NCC approach stress the role of consciousness in learning processes or information integration processes (cf. Cleeremans and Jiménez 2002 ; Crick and Koch 1995 ; Dehaene, Changeux et al. 2006 ; Lamme 2006 ; Tononi and Koch 2008). For example, the distinction between consciousness and pre-conscious or subliminal information processing is claimed to be the accessibility of the relevant information by certain processes (Dehaene, Changeux et al. 2006). And another theory claims that: "the level of consciousness of a physical system is related to the repertoire of causal states (information) available to the system as a whole (integration)" (Tononi and Koch 2008 253, italics in original).

Although Marr referred to an only loose interdependency between the three levels of explanation, as we discussed in section I.3.4, he was aiming for more than just correlations between the theories formulated at those levels. He argued that even though a task could be performed by different algorithms and researchers should therefore allow for potentially different neural implementations, researchers could still use the theories available at a particular level to constrain the number of – theoretically or empirically – plausible theories at another level, and vice versa.

---

[82] Given that NCC research can be distinguished as focusing on either background states of consciousness or on consciousness of specific contents, we would expect that information is a crucial element at least in the latter strand of research. Indeed, information theory is considered to be a major development in he scientific study of consciousness, since "information theory is also the first step in solving the difficult problem of bridging the mental and the physical domains" (Rees and Frith 2007a 14).

Moreover, in this search for mutual constraints he assigned a primary role to the top level or the computational theory (Marr 1982). This is a more specific aim of combining theories at different levels of analysis than the mapping relation involved in NCC research seems to prescribe (Chalmers 2000). Nonetheless, within the NCC we can also discern a fourth strategy: the aim to constrain potential theories on the basis of an overlapping neural correlate that research has yielded.[83] Although there may not have been a very strict delineation of phenomena admitted to NCC research in advance, a post factum comparison of relevant results can suggest a potentially defining neural correlate of those phenomena. Consequently, NCC researchers may be tempted to define the psychological function in terms of this neural correlate and in so doing reverse the lessons that were presented in the first chapters of this part, by the approach that assigned priority to conceptual analysis or Marr's approach. For example, Lamme goes so far as to claim that ultimately, one should redefine consciousness in terms of recurrent processing, since it is the 'key neural ingredient of consciousness' for most – though not all - phenomena associated with the problem of consciousness (Lamme 2006 499).[84] However, this still leaves the demand to explain how this key ingredient can be responsible for the particular phenomena that are associated with the conscious states. For that, a more elaborate specification of the relation between the levels of analysis and integration of such an ingredient with other relevant components is needed than the NCC requires. And eventually, as Chalmers states elsewhere, research of consciousness phenomena cannot stop with the detection of neural correlates but should instead result in: "specifying a *mechanism* that performs the function" (Chalmers 2007 227, italics in original). What such an explanatory mechanism might look like will be discussed in the next section.

---

[83] Here, the mapping relation is not just unidirectional, but bidirectional. Even though the definition mentioned earlier seemed to suggest otherwise, there was also hope that eventually an NCC might help in reaching conclusions about undecided cases like patients in coma or 'locked-in' patients, and that an NCC might help explain the functional phenomena associated with consciousness (Chalmers 2000).

[84] As neurobiological as this key neural ingredient appears to be and as remote from a phenomenological or psychological theory of consciousness, it does bear a similarity with the theory of consciousness at the psychological or computational level which Dennett presents as his concluding hypothesis about consciousness: "our capacity to relive or rekindle contentful events is the most important feature of consciousness – indeed, as close to a defining feature of consciousness as we will ever find" (Dennett 2005 171). Common to both definitions is the fact that contents or information are not just forwarded to further stages of processing.

To sum up this chapter on NCC research, we have seen that even a higher cognitive function like consciousness is open for empirical research aiming to combine different sources of information. In doing so, the NCC approach appeared to be more liberal in certain methodological respects than the other approaches thus far, while some requirements still stand out. Taking the requirement of a strict definition or task analysis of consciousness at first more lightly, the approach still has to use some provisional concept of consciousness in order to decide whether a particular state can be admitted. Furthermore, finding a neural correlate itself is not yet sufficient for explaining its role in the conscious phenomenon in which it is involved. Obviously, words like 'mapping' or 'correlation' that capture the relation between conscious and neural states are mere indices or filler terms that are in need of further specification.[85] What such specification might look like and how these states can be related is left open in the NCC approach and needs other resources for its articulation. Additional resources are also needed in order to acknowledge the differences in relevance between the many neural correlates that will accompany any cognitive function. In the next section, we will discuss such further resources when we focus on the mechanistic explanatory approach. One of the advantages of that approach is that it provides insight into the differences in relevance of the various components of a mechanism responsible for a phenomenon – including the dynamics that underlie the shifts in components' relevance, as their causal relevance may also change under changing environmental conditions.

In conclusion, the NCC approach can at least play a role as a heuristic in commencing research of a complex cognitive phenomenon, inviting research of its neural correlates – in the form of neural localization or otherwise - in the absence of its delineation. However, sooner or later the requirements that were passed over will resurface and ask for fulfillment.

---

[85]  Again, "correlation studies cannot determine whether such neural activity plays a causal role in determining the contents of consciousness" {(Rees and Frith 2007b 560).

# 5  MECHANISTIC EXPLANATION AND THE INTEGRATION OF INSIGHTS*

The three previous methodologies were found to differ in several ways, a major difference being the role assigned to conceptual, empirical or other – algorithmic, for example – insights in the explanation of a cognitive phenomenon. A discussion of these methodologies has among other things yielded the result that pluralism in such an endeavor is inevitable. We noted that defining a complex phenomenon like singing or consciousness will already confront the researcher with likely pluralism. Devising an algorithmic theory – or computational model - for a particular cognitive task similarly offers a plurality of options. Finally, a causal plurality is general involved in a complex phenomenon, to which consequently several distinct theories equally apply, although each with only a limited explanatory power.

A result of this discussion is the need for a methodology that can handle such pluralities, that allows for a phenomenon being given divergent or incomplete definitions, that can integrate different types of theories, and that can handle causal pluralism. The aim of the next and last methodology to be discussed is to present mechanistic explanation as a useful approach that fulfills these desires. Mechanistic explanation requires researchers to determine how their specific explanatory or theoretical insight fits into a so-called 'explanatory mechanism' of a particular phenomenon. A definition for such an explanatory mechanism states that it is: "a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel 2008 13). Fitting an insight into an explanatory mechanism implies first scrutinizing whether it applies to a particular component part or operation, or to an organization feature of the mechanism. Further below we will shed light on how this might work.

The merit of this mechanistic explanatory approach is in our eyes its ecumenical yet not undemanding nature, supporting what Craver calls: "the mosaic unity of neuroscience" (Craver 2007). Notwithstanding its liberal stance with regard to several forms of pluralism, it does require researchers to determine how their specific insights are to be integrated with other available insights in a phenomenon. Another merit is that this approach does not suggest having unlimited applicability. On the contrary, the definition of an explanatory mechanism immediately provides a first

---

* On pages 371, 373, 375 figures I, II, III offer simplified representations related to the arguments made in parts I, II, III respectively. Figure I is particularly relevant as a represention of the main contents of this chapter I.5.

caveat: it is relevant for the explanation of one or more specific phenomena, nothing more and nothing less. We will meet a second caveat later: this methodology does not make the metaphysical or epistemological claim that all phenomena allow such a mechanistic explanation. In fact, there is good reason to assume that there are components in each phenomenon – if only the absolutely smallest ones – that are not explainable in this way.[86] Nonetheless, for the present context and for the explanation of how an agent is capable of 'sculpting the space of actions', this approach offers satisfactory means. In order to appreciate these resources, let us recapitulate some relevant insights from the previous sections.

The last sections devoted to the NCC approach highlighted its liberal use of strategies and heuristics in the investigation and explanation of a phenomenon as elusive as consciousness. Nonetheless, the mere finding of a neural correlate was said only to make sense when it can be interpreted as an ingredient of a more comprehensive mechanism (Chalmers 2007). Other heuristics have been employed by the other approaches. Marr, for example, subscribed to a 'principle of modular design' which implies that "a large computation can be split up and implemented as a collection of parts that are as nearly independent of one another as the overall task allows" (cf. Marr 1976 485; Marr 1982 102). Applying such a principle does not stand in the way of recognizing its limitations, for instance that it "does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular" (Marr 1982 102). In the present, mechanistic explanatory, approach, related principles or assumptions are made concerning the structure of a cognitive process and its explanation, as we will see below.

Another strategy that Marr emphasized was distinguishing between and subsequent integrating 'three lines of attack' (Marr 1982 300) for the analysis and explanation of a cognitive task. In combining three different perspectives – a task analysis, algorithmic theory and neural implementation theory – he invited the interdisciplinary *integration* of several disciplinary perspectives on such a task while allowing the distinct endeavours also some *independence* (Marr 1982).

A third strategy that we discussed above prescribed to first analyze and define the cognitive function or computation under scrutiny. Setting aside their criticism of Marr, Bennett & Hacker agree with him on the importance of such a definition (Bennett and Hacker 2003).[87] As ingredients of such a definitory effort, they specifically mention the use of conceptual analysis and the behavioral criteria that

---

[86] See the discussion of limitations of mechanistic explanation in section I.5.7.

in their view guide our use of the concept of a psychological function. In addition to their view, we argued in section I.2.3 that conceptual variabilities or blurred distinctions between functions (as in blindsight or inattention to pain) can also be used as heuristics, as they sometimes refer to phenomena that are determined by interferences between two different functions or other modulations of a particular function. Our example of singing helped us to spell out such variabilities, underlining the difficulty of providing a single definition of such a function. A requirement of an explanation of such a function is that it helps to provide insight in these variabilities or modifications, which is something that a mechanistic explanation is capable of.

In addition to the lessons drawn from previous approaches, another element needs to be added to the ingredients that a comprehensive explanatory account should be able to integrate. If it is to analyze and explain how development and learning take place, it should be able to accommodate environmental influences, ranging from sensory stimuli to resources that stem from other agents, teachers, etcetera. In contrast to the closed 'behavior program' of the parasitic cowbird, for example, young geese display an open behavior program, as they would continue to associate with humans after being raised by Konrad Lorenz instead of a goose: their imprinting mechanism is to a large extent open for environmental information (Mayr 1964). Indeed, it has been argued that in all dynamic and evolving systems the distinction between innate and acquired traits should be considered gradual and not disjunctive, as environmental information will always become effectively integrated in those systems (Wimsatt 1986). As we will elucidate in Part III how an agent's 'space of actions' can also become constrained by higher order intentions – sometimes stemming from joint deliberation with another agent - we are currently interested in an explanatory approach that permits such influences to determine an explanatory mechanism's behavior.

These requirements and the employment of the strategies that previous approaches have yielded are not easy to fulfil by any account. However, the requirements may gain in plausibility if we illustrate such explanatory efforts with reference to skill learning. Obviously, motor or cognitive skill learning involves many different

[87] Their rejection of applying the notion of 'representation' to anything other than the symbols that humans use to represent things and thus of applying this notion to brain functions (as Marr does) unnecessarily constrains its use and overlooks its *heuristic* use in the study of such brain functions (cf. Bennett and Hacker 2003 70, 76). For recent arguments for and against the use of the concept of representation as part of functional explanations in cognitive neuroscience, see for example (Bechtel 1998 ; Churchland 2002 ; Haselager, de Groot et al. 2003 ; Jacobson 2003 ; Keijzer 2002 ; Piccinini 2008). We will not discuss this issue explicitly, but in Ch. 3 we will argue that verbal representations play a functional role in the mechanism responsible for action determination.

functions, and their complex interactions are modified via dynamic processes like instruction, learning and experience. As a result, automatization of a motor or cognitive skill corresponds to changes in terms of the intentions, consciousness and control involved in it, or in terms of its goal and stimulus dependency, or in terms of its efficiency and speed (Moors and De Houwer 2006). Clearly, an explanation of such a process will be complex and will display intricate dynamic interactions between its many psychological or cognitive ingredients. Similarly, an explanation that focuses less on psychological but more on neuroscientific ingredients will refer to a gradual process involving a complex interaction between cortical and sub-cortical networks during learning and during automatized responses (Ashby, Turner et al. 2010).

Interestingly, such dynamical processes often lead to results that comply with Marr's 'principle of modular design', which was mentioned above. However, what Marr may not have realized is that in many cases modularity is not the starting point but rather the result of a dynamic process, for which the term 'modularization' has even been coined (Karmiloff-Smith 1992).[88] In fact, the mechanistic explanatory approach is particularly well-equipped to yield insight in such a dynamic process, as we will discuss in section I.5.6 below. This fact may to some extent fight Marr's doubt whether dynamic systems allow explanation, believing as he did that his own methodology could not apply to non-modular systems (Marr 1982). Indeed, he held that the Type-2 theories that could explain such complex processes were not available at the time (Marr 1977b). Since then, several such approaches have been developed, especially to cope with such processes, among which the mechanistic explanatory approach which has gained ever more recognition in cognitive neuroscience.[89] If,

---

[88] In his afterword to the new edition of Marr's posthumously published book on vision, his former collaborator Poggio speculates that Marr would have wanted to add to it the process of learning – having been the object of Marr's earlier research – if he had had time (Poggio 2010). One may also speculate how such an expansion of his approach would have required an extension of the methodology, as it is not easily applicable to the complex and interactive systems that demonstrate learning.

[89] Other explanatory approaches that need to be mentioned in this context are those that use Bayesian or probabilistic analysis of processes - like vision, for example (Yuille and Kersten 2006) – or those that commend the use of dynamic systems theory (van Gelder 1998). It can be argued that an explanatory mechanism to a large extent is similar to a dynamic system, depending on whether one assumes that representations play a role in such systems – as was defended by Bechtel (Bechtel 1998) and further refined in (Nielsen 2010). However, dynamic systems theory usually suffices with describing the behavior of a system, not explaining it, which is what a mechanistic explanation has to offer in addition (Kaplan and Bechtel 2011). The role of representation in explanations is by no means a settled issue, nor is the role of representations generally in cognitive neuroscience. First, it should be recognized that representations play different roles (Keijzer 2002). Similarly, the difference between perceptual and conceptual representations deserves to be acknowledged as a way to accommodate critique of representationalism (Markman and Dietrich 2000). An account that seeks a middle road between representationalism and non-representationalism has been suggested, among others, by (Gärdenfors 2004b) who proposes to replace representations with multi-dimensional conceptual spaces – bearing some similarity with the 'space of actions' discussed in this dissertation. (Haselager, de Groot et al. 2003) on the other hand argue generally that the ill-defined notion of representation renders this debate without much use.

moreover, dynamic processes lead to increasing modularization, Marr's hesitations do not need to withhold us from trying to explain such processes.

Let us therefore now discuss how the mechanistic explanation proceeds, integrating the ingredients mentioned in the present section. We will do so by highlighting its three heuristics or steps: definition, decomposition and localization. But first we will give a short introduction of the methodology of mechanistic explanation.

## 5.1 From the mechanicist world picture to the method of mechanistic explanation

Mechanistic explanation aims to explain a phenomenon by providing insight into a mechanism that is responsible for producing it. Whether it is a particular cognitive or behavioral phenomenon, a particular gen expression, or even the warming up of our atmosphere, many phenomena allow mechanistic explanation, thus providing insight into component parts and operations of an organized structure that together produce such a phenomenon (Bechtel 2008). Or, in slightly different terms: "[t]his is a mechanism in the sense that it is a set of entities and activities organized such that they exhibit the phenomenon to be explained" (Craver 2007 5).[90] Below, we will further explain how scientists gather and organize insights into such a mechanism that helps them to explain a particular phenomenon.

The clause that a mechanistic explanation pertains to a particular phenomenon is not unimportant. This specific mechanistic explanatory methodology should be distinguished from the long tradition in the history of science that considers nature as a whole as a mechanism. Indeed, 'mechanicist' world pictures and explanations have been en vogue from antiquity[91] up to modern scientists like Copernicus,

---

[90] There are still other definitions available of the eventual result of a mechanistic explanation – of an explanatory mechanism, that is. As the essential ingredients – a mechanism being constituted by explanatory relevant component parts and operations and their organization – are common to most influential definitions we will not go into the differences (see also definitions in Glennan 2002 ; Machamer, Darden et al. 2000 ; Woodward 2002). While establishing a comparative table of definitions – including some different from those noted here – Hedström comes to a similar conclusion: "Underlying them all is an emphasis on making intelligible the regularities being observed by specifying in detail how they were brought about" (Hedström 2008 321). In another review, he summarizes the telos of mechanistic explanation as: "proper explanations should detail the cogs and wheels of the causal process through which the outcome to be explained was brought about" (Hedström and Ylikoski 2010 50.

[91] An astonishing yet still relatively unknown example of an ancient mechanism is the portable Anthikythera-mechanism, dating back to the second century BC and having no equal within the next millennium. Found by sponge divers around 1900 and first described and partly explained only decades later (Price 1974), the mechanism still surprises researchers, who keep on discovering more astronomical and calendar calculations that it can perform, cf. (Freeth, Bitsakis et al. 2006). We would contend that it also offers a demonstration of the discontinuity of scientific progress, with antique scientists like Archimedes and Aristotle having no immediate descendants until the scientific revolution some two millennia later.

Huygens, Descartes, Laplace, Boyle and Newton and beyond.[92] Inspired by artificial machines or mechanisms, these traditional mechanicist authors tried to reduce all complex phenomena that reality presents to us - usually exclusively material reality, that is – to the products of many, yet simple parts and movements. Most authors recognized that not all parts in nature would allow such a reduction, as it would lead to an infinite regress. However, this did not stop the plea to eventually 'dissolve' all physics in mechanics, as voiced in the slogan that: "all physical appearances have to be explained with reference to natural forces that are exerted by material points upon each other" (Dijksterhuis 1969, translated from Dutch original, p. 538). This metaphysical conviction is specific to this traditional mechanicist world picture.

That is not to deny that there is a methodological interest that contemporary mechanistic explanations share with the traditional mechanicists. Allowing for the importance of the discovery of laws of nature and the probable causal connections behind those laws, both aim to elucidate a mechanism that might be responsible for particular phenomena.[93] However, a crucial difference between the current view on mechanistic explanation and the traditional mechanicist world picture concerns the metaphysical claims that accompanied the latter.[94] Unlike these traditional mechanicist explanations, scientists who currently develop mechanistic explanations do not subscribe to a reductionist and atomist agenda.[95] On the contrary, they do not focus on such metaphysical issues and even "reject any fixed and limited list of the modes by which parts of mechanisms can act and interact" (Glennan 2008 377). This has to do with the recognition that complex mechanisms often have some properties that cannot be explained with reference to such a fixed – and limited - list of parts

---

[92] Obviously, there are important differences between these authors that cannot be articulated in the scope of this dissertation. For example, the implications of Descartes' distinction between extended matter and 'res cogitans' could be relevant. Similarly, Newton's acceptance of gravity as an essential force without an apparent underlying mechanism may be due to his religious perspective, trumping his 'mechanicist' convictions. As for gravity, perhaps a graviton allows its mechanistic explanation, but perhaps the approach is not applicable to this important phenomenon. That would not exclude its applicability to most other phenomena in the material world, though.

[93] Authors defending mechanistic explanation usually contrast this approach with modern deductive-nomological explanation, which implies that an explanation of a particular phenomenon is deducible from at least one natural law. Corresponding with this contrast is the authors' interest in phenomena that do not behave completely law-like and exceptionless (Bechtel and Abrahamsen 2005 ; Craver 2007 ; Machamer, Darden et al. 2000). However, Leuridan warns against focusing solely on explanatory mechanisms while neglecting the observation of regular or lawlike phenomena, since in the life and cognitive sciences such regular phenomena can be multiply realized (Leuridan 2010).

[94] It is therefore not the case that each and every discovery of a causal nexus requires a mechanistic explanation before it can be recognized as such. Indeed, not all discovered causal connections will eventually allow such an explanation, in some cases the 'black box' will remain closed. Besides, it is advisable, for example, to treat an epidemic for which a causal factor has been found, even if no explanatory mechanism has been discovered (Broadbent 2011).

and interactions.

To understand this, it is useful to acknowledge the difference between aggregate, composed and evolving systems. An aggregate system is constituted by a mere collection of parts with simple interactions, such that the addition, subtraction or substitution of parts will not have any effect on the properties of that system. As Wimsatt has argued, this condition holds only for a system's mass, while even the volume of combined volumes of sugar and water for example is not merely aggregative. A system's non-aggregativity is most visible in its having emergent properties (Wimsatt 2007).

In regard to most of their properties, most systems are indeed not just aggregative but composed instead. This implies that to explain most of their properties, scientists must also take into account the organization and interactions that occur between a system's component parts and operations. It is due to such organization and interactions that a composed system displays emergent properties that are irreducible to its smallest components (Bechtel and Richardson 1993). Moreover, a complex system generally displays a hierarchical form of organization in which levels can be distinguished, as this yields advantages in terms of the system's stability and response speed, among others (Simon 1962 ; Wimsatt 2007).[96] This organization structure not only has implications for its interactions with its environment generally, but also for its scientific investigation. Since investigations often involve intervening with a system's properties and detecting the consequences, it will have to take into account the differences between a system's levels and their corresponding properties.

Environmental interactions are not equally important for all systems. Obviously, a sugar solution has a more differentiated response pattern regarding environmental

---

[95] Not all authors subscribing to mechanistic explanation equally shy away from statements with ontological implications. Most explicit in this respect is Wimsatt, who writes about organizational levels in nature, which are occupied by: "families of entities usually of comparable size and dynamical properties" and to which explanations generally refer (Wimsatt 2007 204). Critical of such observations is Craver who denies that levels of an explanatory mechanism are identical to such levels of nature or that the size of the respective entities matters at all. His central point is "that levels of mechanisms are defined componentially within a hierarchically organized mechanism, not by objective kinds identifiable independently of their organization in a mechanism" (Craver 2007 191). Craver's emphasis on this point makes him perhaps oblivious to the intriguing fact that there is a parallel – though not an absolute – between these levels of nature and levels of mechanism, to which Wimsatt draws attention.

[96] Or, to be more precise: heterarchical – a term introduced to refer to a structure of neural networks (McCulloch 1945). In heterarchically organized systems it can occur – due to learning, for instance – that in an explanatory mechanism a particular component at an intermediate level may be bypassed during the performance of a function, compared to the mechanism before such modification (Berntson and Cacioppo 2008). Such heterarchy is observable not just in organisms or the brain, but also in societies, for example, which display more variability and change in power relations than hierarchical structures would allow (Crumley 1995). When reference is made in this dissertation to hierarchically structured and dynamic systems, it is implied that these are in fact heterarchical structures.

changes in temperature, pressure, chemistry, than either sugar or water alone. However, this environmental interactivity or context sensitivity is exponentially greater in systems that can evolve or develop than in aggregate or composed systems.[97]As we will see in this dissertation, the composition of an evolving and developing system is such that it is modifiable as a result of interactions with its environment. In section I.5.6 below, we will emphasize how mechanistic explanation is particularly suitable for the explanation of changes in cognition or behavior due to development and learning, as it can refer to modifications of the explanatory mechanisms. It would be hard or impossible to account for such changes with the ingredients of the traditional mechanicist world picture. For this reason, it offers very limited resources for the analysis and explanation of the regular yet not exceptionless phenomena that pervade the life and social cognitive sciences (Glennan 2008).[98]

Given their suitability for explaining complex and dynamic phenomena, investigations in these fields have often led to the development of explanatory mechanisms, even if not always explicitly so. In their seminal and extensive exposition of the approach, Bechtel & Richardson offer many examples of such research histories: from the Krebs cycle to genomic regulation, from fermentation to cognitive psychology, scientific investigations have resulted in mechanistic explanations while employing the specific heuristics of decomposition and localization (Bechtel and Richardson 1993). More specifically to cognitive neuroscience, similar analyses have been presented for the mechanistic explanation memory (Craver 2002 ; Craver and Darden 2001), for vision (Bechtel 2001b), for action understanding (Keestra 2011 ; Looren de Jong and Schouten 2007), for circadian rhythms (Kaplan and Bechtel 2011), for example. More recently still, mechanistic explanation is being considered as a methodology for the social sciences, with examples referring to social phenomena like the self-fulfilling prophecy or network diffusion (Hedström and Swedberg 1996 ; Hedström and Ylikoski 2010 ; Tilly 2001).

---

[97] The difference in notions of emergence can be generally ascribed to the different – external or internal – contexts with which a system's interactions give rise to emerging new properties (Wimsatt 2006a). For our present purposes, this differentiation is not relevant.

[98] Beatty argues why the converse model – a singular theoretical account of a biological phenomenon – is unlikely and why in a theoretical pluralistic model of explanation there are differences in relative significance between theories (Beatty 1997). Both arguments concur with the mechanistic explanatory approach considered here. 87   With regard to the context of social sciences, it is especially a matter of dispute whether entities such as the state, religion, or collective memory can enter as component parts into an explanatory mechanism, or that these have to be considered only as environmental factors of such a mechanism. Similar questions arise as to the wide variety of social interactions that social scientists include in their explanations of social phenomena: are power or sexual relations fit to be integrated as component operations in a mechanistic explanation of social phenomena? Hedström & Ylikoski, for example, argue for the development of mechanistic explanations in the social sciences with a crucial role for individual agents and their relations (Hedström and Ylikoski 2010).

For each of these subjects, the question as to what component parts and operations are involved in the responsible explanatory mechanism is of primary importance.[99] Moreover, given the non-aggregative nature of such a mechanism, it is likely constituted by parts and operations organized at different levels. Before we move on to a systematic treatment of the methodology of mechanistic explanation via its three heuristics, we will briefly discuss the explanation of memory as an example of this particular method.

## 5.2 Memory and the mechanistic explanation of learning

Since memory and learning are essential phenomena in cognitive science and are also relevant for this dissertation, with its interest in the process of an agent's 'sculpting' the space of his actions, let us consider these as examples for the endeavour of mechanistic explanation. To explain learning accordingly, researchers should provide an ever more comprehensive description of the mechanism that is responsible for it, consisting of component parts and operations and their organized functioning (Bechtel 2008). We must always avoid the assumption that by presenting an explanatory mechanism we are simultaneously providing an exhaustive and exclusive description of all possible functions of the components involved. For especially in complex and dynamic systems that have both an evolutionary and developmental history, it is usually the case that a component is involved in more than just a single function, like it is the cases with a gene that can be co-responsible for several phenotypical properties. Indeed, it has been argued that 're-use' of neural components is a prevalent phenomenon with regard to the brain, implying that many parts and operations figure in more than a single cognitive functions (Anderson 2010).[100]

In the case of memory, cognitive scientists have devoted many investigations to

---

[99] With regard to the context of social sciences, it is especially a matter of dispute whether entities such as the state, religion, or collective memory can enter as component parts into an explanatory mechanism, or that these have to be considered only as environmental factors of such a mechanism. Similar questions arise as to the wide variety of social interactions that social scientists include in their explanations of social phenomena: are power or sexual relations fit to be integrated as component operations in a mechanistic explanation of social phenomena? Hedström & Ylikoski, for example, argue for the development of mechanistic explanations in the social sciences with a crucial role for individual agents and their relations (Hedström and Ylikoski 2010).

[100] Relatively independently and based upon different lines of evidence, several hypotheses have been presented in recent years that are comparable to Anderson's hypothesis regarding extensive neural re-use in the brain (Anderson 2010). For example and building upon evidence about mirror neurons, it has been argued that neurofunctional architecture is being 'exploited' for more than just a single function (Gallese 2008), while cognitive neuroscientific research of reading and writing – which are relatively recent cultural inventions – has suggested that evolutionary older neural circuits are being 'recycled' (Dehaene 2005).

spatial memory in mice. Craver has argued how these studies have culminated in a mechanistic explanation of this phenomenon – and the present section is based upon his presentation in (Craver 2002 ; Craver 2007 ; Craver and Darden 2001). Studying spatial memory, researchers let the animals wander through labyrinths or in the Morris water maze, in which a platform is hidden in opaque water, and measure the time and trajectory they use for finding their way to this platform. Apart from such behavioral measures, researchers can also measure the activities in different brain areas of the animals and try to disentangle areas specifically involved in spatial memory and not responsible for motor behavior or visual processing, for example. Or they can zoom in on a particular area and detect the electrophysiological interactions that occur in the synapses within that area and try to correlate that with particular phases of the animal's behavior. Digging even a level deeper, researchers have investigated the molecular processes that constitute the synaptic electrophysiological
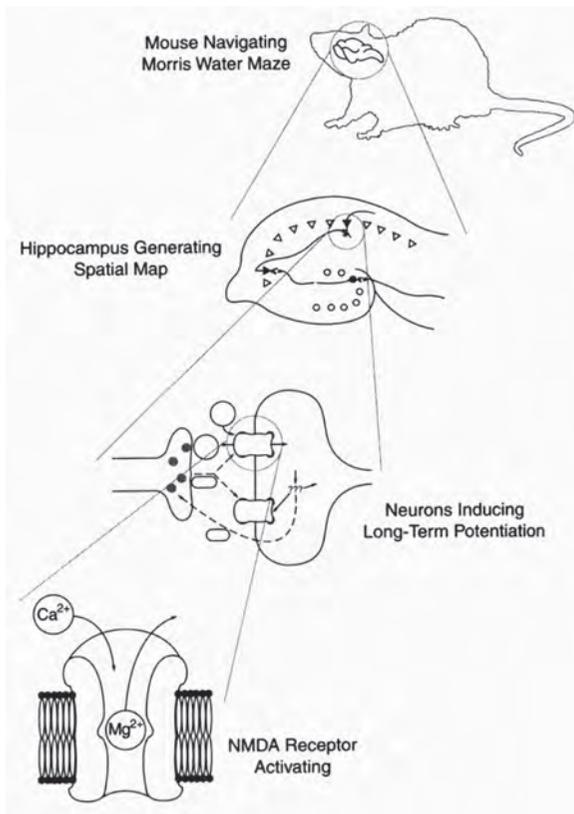


**Figure 1: levels of spatial memory.** Reprinted from (Craver 2007 166) with permission from the publisher

interactions corresponding with the process of Long-Term Potentiation (LTP). This LTP, in turn, occurring in the relevant brain areas under specific environmental conditions, is responsible for the animal's spatial memory learning and the increase of successful navigation. Figure 1 presents a crucial component of the relevant explanatory mechanism of memory, represented at different mechanism levels.

Though we cannot describe the contents of this figure in detail, it may help us to explain what is at stake in mechanistic explanation. Figure 1 comprises figures at four different levels, which all refer to the same phenomenon, namely spatial memory. Spatial memory as such, however, can only be observed as a phenomenon in the mouse's navigating efforts, represented at the top level. Simply put, the main mechanism that was found to be responsible for this memory was located in the animal's hippocampus. So it was this part of the animal rather than other neural or even muscular parts that turned out to be specifically responsible for spatial memory. Obviously, this component part cannot produce successful navigation on its own, as the hippocampal cells still need to interact with neural areas responsible for motor planning, visual perception, and so on. These interactions consist of excitations and inhibitions between neural areas, that together produce the navigation behavior. Research on mice with lesions in the hippocampus has shown that they have specific difficulties in learning to navigate, suggesting a specific function of this neural area.

Knowing that the hippocampus has a specific function for spatial memory does not yet provide insight in the processes that occur in that area. Nor does it shed light on the specific temporal patterns of mice's learning or the sensitivity of this learning process for specific chemical interactions. As Craver argues, the discovery of the sub-mechanism of Long-Term Potentiation as a general synaptic process constituting learning also made it possible to explain specific properties of learning. For example, it turned out that the response – i.e. depolarization - properties of post-synaptic cells correlate with the difference in learning results when an animal is exposed to rapid and repeated stimulation in comparison to single exercises ((cf. figure 5.2.c and text in Craver 2007 168-169). Even though reference to hippocampal activities and their interactions with other areas would be largely sufficient to account for the capability of spatial memory, its temporal peculiarities were therefore better explained by referring to the interactions between singular cells and their changing properties.

These cell properties and interactions are in turn constituted by molecular processes. For the intervention in spatial memory, we can target these processes, if irreversible surgical lesions or difficult electrophysiological lesions have to be avoided. To that effect, we can block the NMDA receptor's channel with particular

chemicals, and thus interfere with the LTP process that is required for successful spatial memory. Here, the component parts are not neural areas or cells, nor are they operations like excitation, inhibition, polarization. At this molecular level we find instead "entities like the NMDA and AMPA receptors, glutamate, Ca2+ ions, and Mg2+ ions engage in activities like attracting and repelling, binding and breaking, phosphorylating and hydrolizing" (Craver 2002 S89-S90).

As a result of mostly separate ethological, neuroanatomical, physiological and molecular research, a comprehensive explanatory mechanism for spatial memory has slowly developed. Part of the research consisted in separate investigations of different processes and parts, sometimes not even directly related to mice navigation skills. Another part of the explanatory efforts involved tying together these separate investigation results into an explanatory mechanism consisting of different levels. This is an important scientific task, as it helps to mutually elucidate the formerly separate results: temporal constraints of spatial learning can be explained with reference to LTP processes, medical intervention can be improved with knowledge of molecular processes, and so on (Craver 2002).

This again confirms the explanatory pluralism mentioned earlier in this part, implying that for a complex phenomenon like spatial memory we invoke – and integrate – theories pertaining to various components of the phenomenon that shed light on its various specific properties.[101] As a result, each theory has only a 'relative significance' (Beatty 1997). The integration of all relevant theories is a difficult part of research and it is here that mechanistic explanation can offer a useful methodology, maintaining as it does the necessary causal and theoretical pluralism.[102]

When providing a methodology for the integration of several theories regarding a phenomenon, it is useful to note that a mechanistic explanation is usually the result of a gradual development. In light of this, a distinction has been proposed between a mechanism sketch and its eventual schema (Machamer, Darden et al. 2000), or between 'how-possibly' and 'how-actually' explanatory mechanisms (Craver 2007). For example, presenting the explanatory mechanism of spatial memory as consisting of three levels of submechanisms, Craver admits that "this explanation into four levels is surely an oversimplification" (Craver 2007 169). Depending on the

---

[101] In other words, integration of different theories amounts to finding constraints at several mechanism levels that together determine the space of options for a phenomenon (Craver 2007).
[102] Wimsatt explains what often happens in a situation where pluralism is required: "given the difficulty of relating this plurality of partial theories and models to one another, they tend to be analyzed in isolation with unrealistic assumptions of system-closure (to outside 'disturbing' forces) and completeness (that the postulated set of variables includes all relevant ones)" (Wimsatt 2007 180). This is similar to the independence and uniqueness that are often and mistakenly ascribed to a classification, although pluralism is usually more appropriate in this context as well (Hacking 1991).

research question and available techniques, the coordinated activation patterns of hippocampal cells or interactions between different neural areas could have been added to the explanatory mechanism, or further inquiries into underlying molecular submechanisms could have been performed.[103]

Finally, for the integration of multiple theories into a comprehensive explanatory mechanism, investigations exploit its multilevel organization and interactions. Interlevel experiments are used both for a bottom-up investigation of a component's function at a higher level – what role does the NMDA receptor play in LTP and consequently in learning? – or for a top-down investigation of the submechanism responsible for a particular function – how does alcohol addictive behavior affect the mechanisms involved in memory formation and learning?[104] In either case, a particular intervention targets specific components at a particular level, while potential effects are detected at another level (Craver 2002). In addition to interlevel experiments 'looking up' for functions or 'looking down' for responsible mechanisms, researchers are also 'looking around' in order to ascertain additional components at the same level or relevant environmental conditions (Bechtel 2009b). In sum, not only does mechanistic explanation provide resources that can help in the integration of various scientific insights, it also assists in articulating modes for further investigating a phenomenon. Moreover, it does so without enforcing a particular metaphysical, reductionist, position.

Now that we have introduced and provided an example of mechanistic explanation, it is finally time to detail the heuristics guiding its development. It is in fact a gradual process facilitated by performing three different heuristics – not just once, but often iteratively. In the following three sections, we will shed light on the heuristics of definition, decomposition and localization. Together, they will clarify how in the next parts the insights from philosophy, psychology and neuroscience can be combined to provide an integrated account of the space of actions that a subject develops, learning to perform his actions in conformation with a wide variety of constraints including his own intentions. To prepare for the explanation of such a dynamic process, the sections below on heuristics will contain references to skill learning and a specific section will be devoted to dynamic modifications of an explanatory mechanism.

---

[103] Indeed, 'bottoming out' an explanatory mechanism at lower levels is relative (Machamer, Darden et al. 2000), and a decision about digging further may be guided even by considerations of its cost-effectiveness (Wimsatt 2007). In the next sections, it will become clear why the explanatory relevance of ever lower levels usually decreases.

[104] Alcohol affects i.a. the NMDA receptor, which we know is involved in LTP (Lovinger 1993).

### 5.3 Defining the phenomenon as a first step

In the context of previous explanatory approaches, we witnessed that definitions of psychological functions or computational tasks generally play a role. However, the role of definitions was quite different. In chapter I.2, we found that Bennett & Hacker argued for the consideration of a function's definition as the a priori delineation of its domain, without empirical, neuroscientific data being able to affect the definition (Bennett and Hacker 2003). Conversely, we found the suggestion stemming from the NCC approach to redefine consciousness in line with the neural correlate of recurrent processing (Lamme 2006). Not content with either approach, we repeatedly referred to the complex phenomenon of singing. By associating singing with similar yet different phenomena, like vocal signalling, infant crying and babbling, bird song, and expert song, we learned that it makes no sense to aim for a single definition of singing in light of these variabilities and dynamics. Since mechanistic explanation is considered to be well-equipped for the explanation of such variable and dynamic functions, the question to be answered is what role it leaves for a definition of the phenomenon under scrutiny.

As expected, defining does play a role in the mechanistic explanatory account as well.[105] As an explanatory mechanism is constituted by component parts and operations that together are responsible for a particular phenomenon, a faulty definition of the latter will have consequences for our ability to develop an adequate explanatory mechanism of it. Such a definition should give us a first delineation of the phenomenon that we are trying to explain. It should also help to limit our explanatory work, since a mechanistic explanation is specifically tied to a particular phenomenon: "boundaries of mechanisms—what is in the mechanism and what is not—are fixed by reference to the phenomenon that the mechanism explains" (Craver 2007 123). A definition helps as a delineation, but it can also misguide research.

Such misguided research can be blamed on several mistakes with respect to this initial step: "because one has tried to explain a fictitious phenomenon, because one has mischaracterized the phenomenon, and because one has characterized the phenomenon to be explained only partially" (Craver 2007 128). As our reflections on singing testify, this admonition to present correct definitions is not as simple as it seems. Plant song may be a fictitious phenomenon, but does that hold for baby song, too? Must we delineate singing from other social or communicative actions, or is that impossible?

---

[105] Strangely enough (Bechtel and Richardson 1993) fails to mention definition as a separate heuristic. This may have to do with the fact that most of their examples do not raise the kind of debate that we find regarding higher cognitive functions.

These reflections also point out the usefulness of the development of a taxonomy or classification as a first step in a research project. In doing so, a phenomenon is being related to neighboring phenomena and separated from others, providing both a first delineation and a first description of the relevant factual domain. As mentioned before, though, it is important to realize that in the life sciences it is generally not possible to provide a single classification for a domain of phenomena, but that pluralism reigns (Dupré 2001 ; Hacking 1991).[106] Even though it is possible to include each phenomenon into a plurality of classifications, such classificatory work can still help to guide further research by setting it into a variety of relations to other phenomena. In doing so, defining can be used as a heuristic, helping even the investigation of phenomena as odd as 'blindsight' (Keestra and Cowley 2011).

Nonetheless, defining and classifying can also go astray and then create specific problems: for example, if spatial memory is in fact an ill-defined phenomenon, it may be doubtful whether its investigations can consistently identify an explanatory mechanism (Sullivan 2010). Most common errors occur when distinct phenomena are erroneously lumped together instead of being assigned to different classes, or when two kindred phenomena are mistakenly split into two different classes (Craver 2007). Clearly, in both cases developing an explanatory mechanism will lead to serious issues: in the first case, it may be impossible to find overlapping components of the mechanism that can simultaneously explain two – actually completely distinct – phenomena. In the second, researchers may find that the explanatory mechanisms of the two phenomena overlap to such an extent that it is worth inquiring which distinctions have kept the two phenomena conceptually separate. In both cases, it may be necessary not only to scrutinize the explanatory mechanism(s), but also to reconsider the definition(s) at stake.

Consequently, defining and delineating a phenomenon and explaining it are intimately related to each other. Although it has been argued in (Bennett and Hacker 2003) that (re-)defining is logically separate from investigating a phenomenon, we hope by now to have convincingly argued that this cannot be maintained, as this position relies on – logically - untenable views of what definitions and explanations are (cf. also Keestra and Cowley 2011). Similar to this recurrent definitory work when developing explanations, it may happen that researchers revisit their initial

---

[106] See for instance the interesting analysis of genos and eidos as relative, not absolute, units in Balme's commentaries in (Aristotle 1972) and related issues in (Gotthelf 1987). The limited value of definitions in Aristotle's biological works is also discussed in (Gotthelf 1997). A modest attempt to relate these issues to mathematical definitions and Aristotle's awareness of the misleading role of language was presented in our (Keestra 1991).

decomposition of a phenomenon and subsequently 'reconstitute' it on the basis of gathered evidence (Bechtel and Richardson 1993).

In closing this section on the relevance of definition for the mechanistic explanatory approach, it may be useful to refer to an example other than bird song, as this dissertation will focus more specifically on human determination of action. We will find in the next chapters that different components constitute the complex mechanism that determines the action a human agent performs in a given situation. Interestingly, this mechanism is also very dynamic, being capable for example of various forms of learning and automatization. If we consider the listed properties of automatization, it already provides a clue for the comprehensive classificatory web that can be developed for defining such automatization. For automatized action has been defined as being largely: "unintentional, uncontrolled/uncontrollable, goal independent, autonomous, purely stimulus driven, unconscious, efficient, and fast" (Moors and De Houwer 2006 297). Obviously, this provides some clues for initial research – which is what a heuristic at least should do. Similar differences can be found between novice and expert song, some components having decreased relevance in the latter. For instance, an expert singer may no longer need to listen to others to keep his tone, or watch a metronome to keep time.[107]

After such a first definition or delineation of the phenomenon, two further heuristics are employed in its empirical research. While definition embeds a phenomenon in a wider web of relations, decomposition then helps to divide the complex phenomenon into smaller portions that facilitate research. Subsequent localization should then help to determine ever more precisely how a phenomenon or its components are performed by a particular organism or system.

## 5.4  Facilitating the explanatory task by decomposing the phenomenon

Notwithstanding corrigibility – or rather pluralism - research generally starts with a definition and delineation of a phenomenon, as we argued in the previous section.[108] Subsequently, its investigation is facilitated by dividing it into explanatory sub-tasks. This is done by applying the second heuristic, that of decomposition. The phenomenon is considered to consist of a number of components, which in their integrated

---

[107] In any case, melodic and temporal processing during singing are largely independent. Relations between specific components of song can differ a lot, including differences between production and perception conditions (Peretz and Zatorre 2005).

[108] In their influential book on 'discovering complexity', the authors present only the two heuristics of decomposition and localization (Bechtel and Richardson 1993). The preliminary delineation of an object appears not to be recognized as a first heuristic, even though it is a task that is different in kind from its subsequent decomposition.

unity amounts to singing, for instance. Remember that an explanatory mechanism consists of an organized structure of component parts and operations which together produce a phenomenon (Bechtel 2008). Decomposing the phenomenon means that several such components are being distinguished, which merit relatively independent investigations and thus facilitate explanatory research.[109]

However, similar to the pluralism pertaining to the classification and definition of a phenomenon, a pluralism of its decompositions is also possible. For instance, in the cognitive sciences phenomenal, mechanistic, functional, and anatomical decompositions are used in parallel for the development of an explanatory mechanism of a particular cognitive function (Bechtel 2008).[110] Obviously, each form of decomposition then focuses on different aspects of it. In case researchers are investigating a cognitive function in order to explain it, it seems plausible to offer at first a phenomenal decomposition.[111] Such a decomposition is usually based upon the observation of behavioral or verbal responses, like when different subjects perform specific tasks under different conditions or when patients are being studied.[112] In the case of memory, this has led to multiple phenomenal decompositions, for example to distinctions between short- and long-term memory, to declarative and procedural memory, and correspondingly to different forms of amnesia (Bechtel 2001a). Similarly and perhaps more familiar are the debates about different phenomenal decompositions of action, starting with Aristotle's notorious distinction between voluntary, non-voluntary and in-voluntary action in his Nicomachean Ethics.[113]

---

[109] Dennett presents decomposition of a computational task also as a way to develop artificial intelligence. Using a nice metaphor, he writes: "[e]ach homunculus in turn is analyzed into smaller homunculi, but, more important, into less clever homunculi.... being reduced to functionaries 'who can be replaced by a machine.'" (Dennett 1978 80). Lacking, however, in this description are the organization of and interactions between the homunculi, without which a decomposed task will not be complete. Note that interactions are also level-specific, with neurochemical interactions taking place at another temporal and spatial scale than the electro-physiological interactions even if the latter are partly composed of the former.

[110] For example, Aristotle was the first to establish a framework for a psychological science and proposing different subdomains or faculties. This has provoked criticism, since his proposed decomposition is held to be outdated (cf. Clark 1997 221; Vanderwolf 1998). The conclusion should be that a specific decomposition may at some point stand in need of correction or even outright rejection as a result of research, but not that we should do away with definition and decomposition as a heuristic entirely.

[111] Wimsatt discusses the descriptive complexity of an organism by aligning several types of its decomposition – e.g. in terms of its physico-chemical or anatomical compositions, its developmental gradients - and proposes a corresponding notion of complexity that refers to the different degrees of non-isomorphism of these decompositions (Wimsatt 2007, cf. figure 9.1, p. 183).

[112] In most cases a phenomenal decomposition, a starter for explanatory research, will be based upon a phenomenal description which is itself akin to a functional analysis: most cognitive phenomena do allow for such a functional analysis. Piccinini and Craver even state somewhat generalizingly: "Psychology should not content itself with the discovery of merely phenomenally adequate functional descriptions that fail to correspond to the structural components to be found in the brain. It should aim to discover mechanisms. To explain in cognitive psychology and neuroscience is to know the mechanisms, and explanation is what functional analysis has always been all about" (Piccini and Craver 2011 308).

A further decomposition, then, is the mechanical one that actually starts to focus on the components that together produce the phenomenon at stake. Although normally hard to distinguish or even perceive in a phenomenal analysis, most cognitive and behavioral phenomena are constituted by a host of component parts and operations. For example, evidence was gathered relatively early on that vision consists of components like color vision, motion vision, face recognition, while two decades ago already more than 30 different mechanism components were distinguished and explained (Felleman and Van Essen 1991). Regarding memory, and irrespective of its phenomenal component (short- or long-term, declarative or procedural, etc.), researchers have traditionally distinguished mechanism components like the encoding, storage and retrieval of memory (Baddeley 1976). Meanwhile, the nature and number of components have been modified, with researchers now seeking to explain additional components like consolidation, re-consolidation, activation and so on (Hardt, Einarsson et al. 2010).

As these examples underline, the results of applying the heuristic of decomposition are likely in need of modification as new insights are gathered. The need for modification will be diminished when the decomposition is the robust result of a variety of studies, and not just of a single type of study - the latter being often the case when a first decomposition of the possible explanatory mechanism is proposed. Exemplifying this is the influential decomposition of language processing, which at first depended heavily on the patient or lesion studies published by Broca and Wernicka. This decomposition suggested a rather simple distinction between speech production and perception (Bechtel 2001c), but has now been superseded by a much more differentiated and nuanced version, although speech production remains elusive partly because animal models are not available for its investigation (Hagoort and Levelt 2009), confirming the need for interdisciplinary efforts.

Even though we have to treat patient or lesion studies with great care, they are often helpful for the first attempt at decomposing a phenomenon's explanatory mechanism.[114] These are then complemented with animal studies, computational studies, experimental studies and so on. Importantly, researchers should not satisfy

---

[113] The difficulty of such decomposition is all the more evident when we realize that Aristotle provided only a twofold decomposition of action in his Eudemian Ethics, rejecting the differentiation of non-voluntary and in-voluntary action. Although Kenny considers the former as an inferior version (Kenny 1979), we beg to differ, as the former decomposition allows an observer to determine the voluntariness of an action even post factum when taking an agent's sorrow or remorse into account ((cf. Ethica Nicomacheia III, 1)

[114] Obviously, extreme care is required when generalizing from pathological studies to the explanation of normal subjects' cognitive and behavioral responses. It is far from clear, for example, whether double dissociations allow us to draw general conclusions, given the complexity and plasticity of the brain (cf. Karmiloff-Smith, Scerif et al. 2003 ; Orden, Pennington et al. 2001 ; Plaut 1995).

themselves with a decomposition based upon only pathological or, conversely, standard conditions. Instead, the decomposition of a phenomenon can be made more robust with the addition of: "descriptions of the multiple features of a phenomenon, of its precipitating, inhibiting, modulating, and nonstandard conditions, and of its by-products" (Craver 2007 128). However, particularly in the case of complex and dynamic phenomena, we may not always get the same decomposition results when investigating different phases of a particular function. This may be the case when we investigate 'nonstandard' cases of a particular phenomenon like in expert singing, or generally in automatized skills. The question is, how mechanistic explanation can deal with such dynamic aspects of a particular phenomenon. We will take up that question in section I.5.6 below.

Particularly with regard to the decomposition of the mechanism, research requires the iterated application of this heuristic, due to the mechanism's structure. As discussed in section I.5.1 and exemplified above with the explanation of spatial memory, an explanatory mechanism is considered to exhibit a hierarchical organization of interacting organized levels. This is not peculiar for mechanistic explanation alone, as Marr also considered such recursive decomposition.[115] In his view, neuroscientific research involved "the study of particular mechanisms, these being assemblies made from basic components" (Marr 1980 199).[116] A similar assumption guides mechanistic explanation, implying that each component of an explanatory mechanism in turn allows further decomposition in terms of its component parts, operations and their organization. Indeed, this assumption concerns the 'near[ly completely] decomposability' of complex systems coupled with their hierarchical organization (Simon 1962). Of course, when employing this heuristic iteratively, researchers must ask themselves whether further 'bottoming out' an explanatory mechanism is still relevant for their explanatory goals (Machamer, Darden et al. 2000) and whether

---

[115] The affinity between Marr and the mechanistic explanatory approach has generally been overlooked in the literature. Bechtel writes, for example, "[e]ntities at different levels of organization stand in a part-whole relation to one another, whereas Marr's levels of understanding involve different perspectives or modes of analysis directed at the same entity or process" (Bechtel 2008 25, n. 11). Craver likes to compare Marr's levels of analysis with the levels of realization in the sense of Kim, while distinguishing these from mechanistic levels (cf. Craver 2007 165). However, given that Kim considers his levels also along the mereological lines just like Craver does (Kim 2000), Marr may be disagreeing less with both Bechtel and Craver than they assume.

[116] Although this (fourth) level may not be a principal addition to his methodology – reason for its relative absence in Marr's writings – it does for us signal two relevant aspects. First, it emphasizes that Marr's methodology refers not just to the relation between different theoretical perspectives, but also to the nature of the object of cognitive science. Second, by proposing these two neural implementation levels, Marr also appears to assume that recursive decomposition is a necessary ingredient of research, suggesting to us a particular structure of the explanatory mechanism that may be its result.

continually expanding their research in this way is cost-effective (Wimsatt 2007).

In a converse direction and less common in cognitive neuroscience, a phenomenon can be taken as an non-isolated explanandum ready for decomposition. Instead, often a phenomenon also figures itself as a mere component in an overarching and complex phenomenon. In such a case, researchers are looking for the 'role function' of the original phenomenon in its wider context, which usually also helps to explain several of its properties (Craver 2001). This applies when researchers aim to account for the role function of spatial memory in mouse navigation, of singing or skill learning. However, not every single functional interaction between a particular phenomenon and other phenomena is a satisfactory basis for assuming that they are in fact components of a distinct and overarching explanatory mechanism. If spatial memory and navigation only co-occurred at rare and exceptional moments, for example, their being a part of a comprehensive mechanism would seem unwarranted.[117] However, if they always co-occur, neglecting spatial memory's role function likely impedes its explanation. These considerations will return in the next section when the third and last heuristic, localization of an explanatory mechanism, is at issue.

## 5.5  Localization of the decomposed phenomenon

After defining a phenomenon and subsequently decomposing it into components, researchers aim to find a 'locus of control' for the responsible mechanism or one of its components by tentatively localizing it somewhere in the system or organism that displays it (Bechtel and Richardson 1993). Localization in itself being common to scientific efforts, the mechanistic explanatory approach can help to further clarify its procedures.[118] Such localization efforts are meant to further determine the plausible options for a mechanism that is responsible for a phenomenon and to exclude regions from this 'space of possible mechanisms' (Craver 2007 247-248). Before we describe how such localization may be carried out, a reservation has to be made.

---

[117] This naturally raises the question whether this explanatory approach accepts the thesis of an extended mind (Clark and Chalmers 1998), of embedded or distributed (Hutchins 2010), or enactive (Di Paolo, Rohde et al. 2010) cognition. Generally, this approach is not unsympathetic towards it, allowing mental mechanisms to perform functional roles in overarching phenomena. However, as soon as the locus of control has to be identified, and spatial, temporal or processing constraints are to be specified, it will often turn out that an individual agent is relatively autonomous vis-à-vis his environment (Bechtel 2009a). It is doubtful whether the 'transparency constraint' suggested by Thompson is strict enough and complying with these other constraints (Thompson and Stapleton 2009). Indeed, claims for embedding or extending mental mechanisms can easily be overstated.

[118] Although one may wonder how it is possible that, after acknowledging the complexity of a function and its emergent properties researchers still believe they can localize a complex function in an organism, Wimsatt rightly notes that it is a: "howling non sequitur that functional organization is not physical" (Wimsatt 2007 190).

As one of the most thought-provoking phenomena in contemporary physics is the appearance of non-locality or non-localizability, assigning a prominent role to the heuristic of localization may cause wonder.[119] Similarly, many large-scale systems or dynamic interactions between systems also appear to withstand localization, if localization is taken to imply the reduction of systemic properties to the properties of their distinct constituting entities. Three responses to this reservation are in order. First, as declared above, mechanistic explanation is useful to integrate and organize explanatory insights that are of a pluralist nature. Yet it does not have the ambition to replace all other forms of explanation, as it is specifically fit for 'near decomposable' phenomena (Bechtel and Richardson 1993 ; Callebaut and Rasskin-Gutman 2005 ; Wimsatt 2007). Second, not only are many phenomena not 'near decomposable', even for phenomena that are decomposable it is most likely that at some level of decomposition we will meet components or operations that resist further decomposition and thus resist further mechanistic explanation. However, the lowest level of a particular explanatory mechanism is *not* even meant to be the same as a lowest level of reality (Bechtel 2008 ; Craver 2007).[120] Third, even if such non-decomposable components are found at the bottom of an explanatory mechanism, this does not imply that such explanation is reductionist in the common sense of the word. Given its emphasis on the organization and interaction between components, which add additional levels of new, emergent properties to the mechanism, the acknowledgement of non-decomposable and non-localizable entities does not contradict their involvement in phenomena that do allow mechanistic explanation (Wimsatt 2007).[121] Indeed, almost every phenomenon in the material world is a demonstration of this obvious yet often confused or misunderstood fact. Leaving this misunderstanding behind, let us now look closer at the heuristic of localization.

---

[119] Thought-provoking as the might be, the interpretation of the modal status of these appearances is far from settled. Indeed, Dieks argues that an empiricist, Humean interpretation has much its favor (Dieks 2011).

[120] This is why reductionism within mechanistic explanation is qualitatively different from reductionism in the common sense. In Bechtel's words: "if we adopt the mechanistic account, in which the notion of levels is defined only locally, then we are not confronted with the prospect of a comprehensive lower level that is causally complete and closed" (Bechtel 2008 148). Moreover, as noted earlier, mechanism levels are relevant for the explanation of a particular phenomenon and have no general ontological status, in contrast to the levels implied by common reductionist views.

[121] In neuroscientific terms, the crucial role for organization and interaction implies that not only neural cells but also their connections matter when it comes to decomposition localization – grey and white matter both matter (Ross 2010). Techniques for the precise imaging of connectivity in the living brain have become available only recently, so insight is still limited. One proof of its relevance is evidence that disturbed – anatomical or functional - connectivity may be partly responsible for cognitive dysfunctions (Andrews-Hanna, Snyder et al. 2007 ; Courchesne and Pierce 2005 ; Minshew, Williams et al. 2009).

Localization of cognitive functions has been a common strategy since the prehistoric days of trepanation, when skulls were penetrated probably to alleviate pathologies like brain haemorrhage (Missios 2007 ; Verano and Finger 2009). In modern times, localization was employed by phrenologists like Lavater and Gall to directly associate neural areas with psychological faculties or capabilities, influencing the investigations of Broca and Wernicke. The general assumption that even complex cognitive functions are localized, in an undecomposed fashion at a single location in the brain, was already matter of debate in the 19th century (Barker 1995) and has by now been largely abandoned. Still, remnants of such simple localizations still pervade the neuroscientific literature and are being critiqued as the 'new phrenology' (Uttal 2001). However, when researchers aim to develop a mechanistic explanation for a particular cognitive phenomenon's components, a direct correspondence between such a comprehensive phenomenon and a particular brain area is not at issue (Bechtel 2002).[122] Moreover, as we will see, the use of localization as a heuristic can be valuable even if in many cases it will have only limited success.

Whether it is on the basis of only a phenomenal decomposition, or based upon a first decomposition of the responsible mechanism, localization involves the further investigation of a phenomenon's physical properties. In cognitive neuroscience, this implies the formulation of heuristic 'psycho-neural identities' (McCauley and Bechtel 2001). As the notion suggests, localization here implies both the heuristic identification of two different levels of analysis – here: psychological and neuroscientific levels – and simultaneously also the distinction between levels of the explanatory mechanism.[123] This is comparable with Marr's approach, who admitted a loose relation between his computational and neural implementation theories of a particular function. Moreover, Marr also recognized that the investigation of a function's implementation in a particular mechanism must be directed at several levels, as it consists of components that are assembled in an organized fashion (Marr 1980). This was evident from our example of the spatial memory of a navigating mouse, which was localized somewhere in its hippocampus, while subsequent research further localized relevant components present in that area.

Often, deciding between different options of a phenomenon's locus of control and

---

[122] This observation is also relevant for the discussion on double dissociation studies in neuroscience, cf. footnote 114 above.

[123] Compared with Marr's reservations against strict interdependencies between his computational and implementation theories (Marr 1982), mechanistic explanations explicitly aim to explore the potential constraints available between different kinds of levels in order to determine explanations. In so doing, the levels of processing, organization and analysis as distinguished in (Churchland and Sejnowski 1988) are also used in combination and not only separately.

about its further components involves an element of choice, without the availability of strict criteria for making that choice. Wimsatt has proposed to consider the robustness of a phenomenon or a component as a criterion. The more robust a phenomenon is, he argues, the more easily detectable it will be by independent investigatory methods, and hence the more explanatorily fruitful and predictively richer (Wimsatt 2007 63). Clearly, opinions can diverge on the degree of robustness or relevance of a particular phenomenon or component thereof, as we will see. Indeed, controversies in the field of cognitive neuroscience often depend on such divergences.[124] However, with growing evidence, researchers can usually localize a phenomenon or its components ever more precisely.

An initial step in such a localization effort is to 'segment a system from its environment', for example by investigating whether the systems displays the phenomenon in different environments or under variable environmental conditions (Bechtel and Richardson 1993). Spatial constraints are indeed among the most relevant constraints that help limit the space of possible explanatory mechanisms for a particular phenomenon (Craver 2007). This usually first involves a designation of its locus of control as a whole, with later refinement as information about the spatial constraints of its components and their organization are obtained. Similarly, temporal constraints concerning the order, rate, duration and frequency of relevant activities play a role in determining the plausibility of a phenomenon's locus of control, particularly when considered in parallel with the spatial constraints. For example, at what particular locations are activities observable, during which phase of a particular phenomenon (Craver 2007)?

If researchers agree on a particular (sub-)system as a locus of control for the performance of a phenomenon independent from its wider environment, subsequent research may then seek to localize it – or its components - ever more precisely in a particular part of it.[125] Just like the recursive decomposition of a phenomenon implies its having a hierarchical structure, most localization techniques in cognitive neuroscience assume that a phenomenon is produced by a hierarchically structured

---

[124] In his critique of mechanist explanation, Moss focuses particularly on Craver's work and its normative tenor (Moss 2012). Indeed does Craver not seem to realize the consequences from the non-rigid nature of the robustness criteria mentioned here. On the other hand, Craver does emphasize the limited nature of any explanatory mechanism, valid as it is only for a particular phenomenon (Craver 2007).

[125] In cognitive neuroscience, even such segmentation is often disputed. Indeed, currently it is much debated whether cognition should be considered to be not only 'embrained', but also 'embodied' or 'situated' in a broader sense (cf. discussions in (Anderson 2003 ; Barsalou 2010 ; Mareschal, Johnson et al. 2007 ; Niedenthal, Barsalou et al. 2005 ; Wilson 2002). Where embodiment will generally satisfy some spatial constraints as the body 'travels along' with a brain, this will be much more debatable for broader interpretations of situated cognition. The connection between a cognitive phenomenon and an environmental condition is much looser, obviously, which is reason for caution with regard to further assuming cognition's structural embeddedness in more comprehensive systems.

mechanism. Above in section I.5.2, we mentioned the three 'directions' of research appropriate for the multi-level structure of explanatory mechanisms, as investigations amount to looking down, up and around (Bechtel 2009b). We referred to 'inter-level experiments', that is: interference, stimulation and activation experiments. Such experiments intervene at a certain level of the mechanism and aim to detect the consequences of the intervention at another level – either top-down or bottom-up.[126] For example, spatial memory was localized in the mouse's brain, with components being determined at several levels: hippocampal cells, synaptic processes involved in LTP, NMDA receptors (Craver 2002, see the fig. in the previous section ).

These inter-level interventions and detections must involve different methods according to the level at which they aim, because levels are occupied with different component parts and operations and because new properties emerge at each level due to the organized interactions of the components at lower levels (Bechtel 2008).[127] Given these level-specific properties, we can employ stimulus-response studies, imaging studies, single-cell recordings or pharmacological studies in order to detect memory responses, for example, or how singing behavior alters due to such local interventions. The results will pertain to localization at different levels, showing the involvement of brain networks, brain areas, particular cells, and specific molecules. Consequently, researchers aim to integrate not only spatial and local constraints but also other relevant constraints when determining a particular phenomenon's explanatory mechanism (Craver 2007).

Similar to its decomposition, the search for the locus of control of a cognitive phenomenon often also benefits from brain lesion studies in patients. For localization purposes, this technique is often not very reliable given the brain's plasticity, evident in reorganizations that occur in a disrupted yet dynamic neural mechanism (Buonomano and Merzenich 1998). Nonetheless, Broca's and Wernicke's

---

[126] As mentioned earlier, it is important to note that in a complex and dynamic mechanism, other than in a purely aggregative system, there will be many properties that emerge at higher levels. Such emergent properties may in turn have interactions with lower level properties of the mechanism and also allow forms of interaction with the environment that the lower level components and operations by themselves are incapable of (Wimsatt 2007, particularly part III).

[127] Investigative techniques are often specific depending upon the mechanistic level at which they are applied. Wimsatt explains this by referring to organizational levels in nature, which are occupied by "families of entities usually of comparable size and dynamic properties, which characteristically interact primarily with one another, and which, taken together, give an apparent rough closure over a range of phenomena and regularities" (Wimsatt 2007 204). Generally, therefore, reduction is not plausible, and even less so in the case of a specific explanatory mechanism, where levels differ as a result of the organization of components at each level, as Wimsatt argued already in his (Wimsatt 1976). Successful reductionism of theories that apply to a particular level is also much rarer than its notoriety suggests. With regard to interlevel theory reduction, McCauley concluded even that: "The history of science reveals no precedent for theory replacement or elimination in interlevel contexts" (McCauley 1986 197).

studies of lesion patients suggested a particular decomposition of human speech, but additionally allowed these neuroscientists to provide a first localization of specific speech components in particular brain areas.[128] However, as mentioned in the previous section, speech has meanwhile been recomposed or reconstituted on the basis of subsequent research, corresponding with different and more elaborate localizations of the explanatory mechanism for speech (Bechtel 2001c).

Indeed, as inevitable and important as localization efforts in cognitive science may be, localization hypotheses are likely to require revision continuously. Such revision usually corresponds with more detailed insights in the constraints of the phenomenon itself and its explanatory mechanism. An extra complication for localization efforts is the fact that cognitive phenomena are also malleable through development and learning. This is the case with spatial memory, as it is with skill learning, expert singing. The question is whether or not such learning modifies the mechanism that is responsible for such a phenomenon. And if the explanatory mechanism changes during learning, is its localization different for novices and experts? Clearly, we are once again facing the question whether novice and expert performance can be explained with reference to a single explanatory mechanism, or whether we should distinguish between mechanisms responsible for their performance. This issue will return in the final section, with some concluding reflections on mechanistic explanation. But first, we will consider how mechanistic explanation can accommodate such dynamic modifications.

### 5.6  Mechanistic explanation and mechanism modifying dynamics

In what follows below, we will try to sketch some forms of mechanistically explaining the alterations involved, for example, in learning a particular skill. Even though it is generally acknowledged that our mental mechanisms "are plastic mechanisms that develop and change as a result of experience" (Bechtel 2008 240) and that learning in the sense of Long Term Potentiation has been explained mechanistically (Craver 2007), a more systematic treatment of such development and learning in the common sense has not been provided in the literature on mechanistic explanation, as far as we know.[129] Such a treatment would involve the articulation of various forms of

---

[128] Some other examples of decomposition and localization efforts in cognitive neuroscience stem from the research of memory (Craver 2002 ; Craver and Darden 2001), vision (Bechtel 2001b), and action understanding (Keestra 2011 ; Looren de Jong and Schouten 2007). Bechtel considers vision research as an exemplar in the Kuhnian sense, or a model, for the development of mechanistic explanation in cognitive neuroscience (Bechtel 2001b).

modification of an explanatory mechanism under the influence of experience or learning processes. Below, we will list four different forms of modification and present some empirical illustrations.

But first, a more principal remark is in order. Obviously, not all modifications in a multi-level mechanism are equally relevant. For example, distinguishable dynamic changes at lower levels of a multi-level mechanism can occur with high frequency without affecting relevant changes at higher levels, like changes in behavior.[130] This has to do with the 'dynamic autonomy' of mechanistic levels, entailing that most "micro-level changes don't make a causal difference at the macro-level" of a system (Wimsatt 2007 218). Such dynamic autonomy is a challenge to reductionists who believe that all events at the lowest levels transpire into appearances at higher levels of a mechanism and that the latter have no relative independence.[131] Conversely, modifications at higher levels of a system may not always be easily detectable as well.[132] In this dissertation, for example, we will discuss the dynamic changes that take place in the mechanisms involved in human action determination. In some cases, development leads to a reduction of the complexity of this process that is not always detectable in human behavior. For now, let us continue the exploration of the dynamics that can modify an explanatory mechanism and subsequently discuss the role of component parts and operations, their organization, and the mechanism's interaction with its environment.

A first modification to be considered is related to the set of mechanism parts that

---

[129] Bechtel has in some recent publications elaborated on the inclusion of dynamic systems theory in mechanistic explanations of dynamic functions like circadian rhythms (Bechtel and Abrahamsen 2010), arguing that the two methodologies can complement each other (Kaplan and Bechtel 2011). Although reference will be made to this work here, it has not really touched upon the kind of processes that interest us, nor has it systematically investigated the possible modifications of an explanatory mechanism due to dynamic changes.

[130] As mentioned in footnote 48, Wimsatt makes this point regarding multiple realizability and notes that philosophers of psychology tend to overlook the prevalence of such multiple realizability (Wimsatt 2007). However, Bechtel warns that many observations of multiple realizability of cognitive functions fail in drawing an adequate comparison between the different instantiations of a phenomenon and particularly their functional characterizations (Bechtel and Mundale 1999).

[131] This dynamic autonomy of – particularly higher - mechanistic levels is partly due to the fact that organization forms in complex systems usually are relatively robust, making them less vulnerable to disruptions. At least two different forms of robustness organization can be found in evolved, developing systems: redundancy robustness – with several identical pathways in parallel - and distributed robustness – with a network of non-identical, alternative pathways (Felix and Wagner 2006).

[132] Simon notes in a similar argument that most systems are near-decomposable, making for an incomplete dynamic autonomy of their levels. Hence, our account of these systems will always: "fall short of exactness because the properties of the lower-level, higher-frequency subsystems will 'show through' faintly into the behavior of the higher-level, lower-frequency systems" (Simon 1973 25). As a result of this, the behavior of such systems are not so much law-like, but rather in the form of regularities that include some exceptions (Glennan 2008).

are involved in the explanandum phenomenon. Remember that we are not talking about the introduction of a new part into the organism or its brain – only the novel involvement of an available part into the explanatory mechanism for a phenomenon. Given the fact that neural 're-use' is common in the brain, we may expect alterations with respect to a mechanism's components as well (Anderson 2010). Particularly interesting for our purposes is the addition or deduction of components due to learning or experience, for example when LTP creates strong interactions between previously loose neural areas. Alternatively, a modular component can even emerge over time due to plasticity, when internal connections within a particular neural network are strengthened above a certain threshold.[133] Again, such a modification can develop while potentially leaving the phenomenon largely – particularly in standard conditions – intact.

The second modification will involve the operations that are performed by component parts of a mechanism. Again, LTP with its activity dependent alterations in synaptic responses and associated genetic, neurochemical and molecular processes (Bliss and Collingridge 1993) may figure here as an example of modified activation patterns. Just like learning in a subject often involves the modification of behavioral or verbal responses in specific situations, such learning is often constituted by dynamic changes of operations performed at lower levels of the explanatory mechanism. Depending on the rest of the mechanism involved, such a modification can influence a cascade of further mechanism activities, leading to rather novel behavior of the mechanism. However, the results may also be more modest, as when enhanced specific stimulus sensitivity of a component merely leads to increases of the speed and efficiency of a subject's responses.

Third, the organization of the components may be modified via development or learning. As mentioned earlier in these sections on mechanistic explanation, it is the organization or re-organization of components that is often responsible for the emergence of new properties within a mechanism. In a hierarchically – or rather: heterarchically[134] - organized mechanism, learning and experience can affect the configuration of the component parts and their interactions, thus altering a phenomenon. It can involve, for instance, the alteration or even the thinning out of the relevant organization, an intermediate mechanism component being skipped when

---

[133] Such 'modularization' as a result of learning and development will be treated more extensively in chapter I.2, which discusses neuroconstructivist accounts like those presented in (Karmiloff-Smith 1992 ; Mareschal, Johnson et al. 2007).

[134] See (Berntson and Cacioppo 2008 ; McCulloch 1945) and footnote 96 on the importance and prevalence of heterarchy.

direct connections between more distant components have developed. Alternatively, feed-back loops can develop, or two previously unconnected networks can become dynamically coupled, making the organization much more complex than in a serially organized or linear mechanism (Bechtel and Abrahamsen 2010). In this context, too, such complex organizational modifications will more likely obtain at lower levels of an explanatory mechanism and will not be directly mapped onto identical changes in the phenomenon to be explained.

A fourth and final mechanism modification that can occur is of a somewhat different character, as it involves more than just its internal components or organization. One of the results of increasing complexity of a mechanism and the emergence of new properties at higher levels is its expanding capability of interactions with the environment. Together with the increasing degrees of freedom that such a system owes to the emergence of properties at higher levels, there comes an increase in such interaction capabilities, as with the development of molecular compounds, with sensory systems, with locomotion, and so on. Put in simple words: "There should be more ways of interacting with a spouse than with a quark!" (Wimsatt 2007 223). Given plasticity and learning capabilities of neural mechanisms, these expanded interaction capabilities can also have a lasting impact on the internal composition of the mechanism.

It is important to note, however, that a mechanism's development and learning does not always lead to an increase of interactions. On the contrary, these processes can also yield new strategies for complexity reduction. This will be a topic in the next parts, where the process of 'sculpting the space of actions' will to some extent consist of such a reduction, as it involves an increased consistency and coherence of action – a welcome phenomenon when interacting with a spouse or in singing, for example. Similarly, learning often results in complexity reduction by reducing the number of dimensions of a content through foregrounding some of its dimensions to the detriment of other dimensions, for instance by chunking memorized contents – as will also be discussed in the next part (cf. Halford, Wilson et al. 1998 and commentaries).

These four modifications are the most prominent ones that can affect a mechanism performing a particular phenomenon. A few empirical examples may help to clarify these modification types. A prominent phenomenon discussed in this context is that of modularity, which is related to the modification of component parts and operations within an explanatory mechanism.[135] Generally, modules are considered to be – functional, if not anatomical - components which have some autonomy within a system or mechanism, performing a specific function, sometimes related

to a specific information domain (Barrett and Kurzban 2006 ; Mitchell 2006 ; Seok 2006). Interestingly, such modules often appear as the result of a process in which a mechanistic component develops through increasingly interacting neural networks, which as a result become increasingly specialized in specific operations and inputs (Karmiloff-Smith 1992). Such a modification also affects further operations within the mechanism, since increased interactions within a modular configuration of mechanistic components are usually associated with decreased interactions with external components – like with other brain components or with the environment (Meunier, Lambiotte et al. 2010).

Generally, therefore, we may say that in many cases a combination of different modification types will occur, as in the case of skill learning, where several modifications occur in parallel. fMRI studies of skill automaticity suggest that automatization relies to some extent on increasingly efficient neural interactions within an existing network, based upon a form of Hebbian learning. Such learning alone would not modify the explanatory mechanism in a far-reaching sense. Indeed, an additional and different type of change has been observed, involving the alteration of the neural areas that are being recruited during skill performance (Petersen, van Mier et al. 1998). This is due to the fact that automaticity of a particular cognitive or behavioral phenomenon in many cases involves a shift from deliberate action planning to direct stimulus-response associations (Graybiel 2008).

So it is not just an increasing efficiency of some operations at neural levels of the mechanism that is observable in experts, but also a recruitment of different neural areas, affecting higher levels of the explanatory mechanism. As usual in such complex mechanisms, other modifications are also observable – even if not directly relevant. Depending on the modifications, extra neural resources become available to experts. Consequently, experts can more easily perform an additional task without the skilled task being disturbed, while novices in such a case must exert extra control and recruit extra frontal areas (Poldrack, Sabb et al. 2005). The net result of these changes is therefore not just increasing speed and efficiency pertaining to the skill, but also an increasing capability of responding to other internal or environmental conditions, enhancing the flexibility of the expert. An expert singer, for example, is flexible in meeting the intonation problems of an accompanying instrument or a conductor's forgetfulness, while a novice singer is not.

As a result of these observations, we may well conclude with regard to complex

---

[135] In the next part, modularity will be discussed more generally. Let it be noted here that since (Fodor 1983), modularity has been much debated and has received many different interpretations, cf. the review in (Seok 2006).

and dynamic systems that apparent identity in cognition or behavior may well hide differences in the relevant explanatory mechanisms. Differences that are often undetectable when applying standard criteria but that can at other times play out and therefore raise questions. It is long overdue, for that matter, that the impact of cultural differences on brain and cognitive processes and their large-scale neglect in cognitive neuroscience is addressed (Arnett 2008 ; Henrich, Heine et al. 2010). An important issue would be whether enculturated brains differ only in their functional anatomy or perhaps even in their structural anatomy, making these differences more resistant to change (Han and Northoff 2008). For mechanisms in general tend to restabilize after having undergone modifications, especially as subsequent actions and developments build on these, as we will discuss in the next part (cf. Wimsatt 2007). It is therefore relevant to note that functional effects of culture-specific differences have been detected even in relatively 'simple' cognitive processes like perceptual information processing or direction of attention (Nisbett, Peng et al. 2001). Given the importance of perception and attention for environmental interactions, development and learning, the effects of such differences could be pervasive.[136] After all, whether it is for the learning of skills, memorizing of information, or even singing, perception and attention exert influence on these processes.

With these considerations of mechanism modifying dynamics, our long exposition of the mechanistic explanatory approach has nearly come to an end. Even though this approach combines several of the merits of the other approaches while adding some more advantages, in the next section we will observe that it has also some limitations, a number of which are neither new nor specific to it.

### 5.7  Some limitations of the mechanistic explanatory approach

After having presented the mechanistic explanatory approach rather extensively, let us not overlook some limitations or reservations that cling to this approach as well.

First, it was noted that this approach cannot escape from the difficult problem concerning the identity of a phenomenon and of its explanatory mechanism. Obviously, this should not surprise us, as this problem was already a concern in Aristotle's reflections on scientific explanation (Sorabji 1980).[137] As is evident, the problem will return at several phases during the development of a mechanistic explanation and not just surface with regard to the phenomenon's definition.

---

[136] Our earlier discussion of the distraction of attention from pain in section I.2.4 has taught us that even pain processing is sensitive to attention's functional role. As another example, depending on the attended feature of a set of stimuli, subjects will adapt their categorization of objects which will subsequently affect their interactions with these objects (Blair, Watson et al. 2009).

Indeed, it is the interdependence between all three heuristics and equally the interdependence between mechanism levels that can help researchers to cope with this problem. Forming a definition helps to decompose the phenomenon, mechanical decomposition may then receive confirmation or disconfirmation from preliminary localization efforts, which may have us revisit the definition, and so on. In the end, it is its overall robustness that supports a particular explanatory mechanism (Wimsatt 2007).

Second, it must always be kept in mind that an explanatory mechanism is related to a particular phenomenon and not a complete description of a system or organism that performs that phenomenon. This also implies that we have to be extremely cautious when generalizing the insights into a particular mechanism to another, neighboring, phenomenon. The constitution of a responsible mechanism for a cognitive function is not equal to the general constitution of the brain as particular components may operate in a different configuration when involved in another cognitive function, or: "levels of mechanisms are defined componentially within a hierarchically organized mechanism, not by objective kinds identifiable independently of their organization in a mechanism" (Craver 2007 191). This lack of generalizability of mechanistic insights is not equal for all levels, of course. The molecular interactions found in NMDA receptors are likely present in all such receptors in the brain, whereas the properties of hippocampal cells are different from other cells in that area: insights in components at higher mechanistic levels are mostly less generalizable, since those components have smaller prevalence and appear with greater diversity. The number of components in quantum theory is small, yet notwithstanding the huge numbers in which they appear, they always behave according to general principles. It is hard to maintain this for the relatively small number of individuals that make up global society.

A third reservation concerns the 'near decomposability' that is assumed by explanations that refer to multi-level systems. Simon did acknowledge that this assumption has both an ontological and an epistemological side to it: when systems are not decomposable, it will be hard if not impossible to explain their behaviors (Simon 1962). A similar position was taken by Marr, who also assumed a system to have a modular and hierarchical organization (Marr 1982). In the previous section, however, we saw that in some cases increasing decomposability is a matter of

---

[137] Indeed, since Aristotle it is a major problem how to present a being's integrated unity in its definition, which instead focuses on the being's distinctive properties, cf. (Kessler 1976). It is important to realize that Aristotle's interest in this issue is especially motivated by his interest in biological kinds and much less by an interest in Platonic, geometrical objects, even though many interpreters have overlooked this fact (Aristotle 1972).

development and learning. Indeed, when a neural network is not yet decomposable and localizable, its performance may be rather difficult to explain mechanistically (Bechtel and Richardson 1993). Generally, in cases where a system's modular and hierarchical organization breaks down and organizational homogeneity increases, when consequently levels are hard to distinguish and internal regularity decreases, we may need yet another methodology (cf. Wimsatt 2007 221 ff.).

This leads naturally to the fourth observation, that in any case at the lowest levels of each mechanism, recursive decomposition and localization come to an end. As was noticed above, our explanatory goals in cognitive science are mostly satisfied with insights in mechanistic levels above subatomic levels.[138] Delving deeper, researchers will inevitably reach components that are no longer decomposable. This may also hold for an omnipresent force like gravity, for which probably no explanatory mechanism can be presented. This limitation of mechanistic explanation with regard to components at a potential fundamental level does not hinder its applicability to most other phenomena.[139]

Finally, it must be emphasized that mechanistic explanation may have been developed as an alternative to deductive-nomological explanations, but law-like theories will still figure within a mechanistic explanation at many places.[140] More than anything else, it is the ability of the mechanistic approach to integrate interdisciplinary results of research that makes it suitable for the demands of cognitive science (Keestra 2011). Indeed, given our conviction that the explanation of human action must allow room for a causal and theoretical pluralism, mechanistic explanation's explicit acceptance of such pluralism is the reason for adopting it as a leading model of explanation in the following parts. Once again, this does not imply that all phenomena pertaining to human action allow mechanistic explanation.

---

[138] There have been several attempts at an explanation of consciousness – which for some has a surprising indeterministic aspect – with reference to quantum physical processes that take place in the brain (Hameroff and Penrose 1996 ; Koch and Hepp 2006 ; Libet 2004 ; Walter 2001). However, not only is it implausible that the specific properties of consciousness can be tied to the myriads of quantum phenomena in brain cells, it is also unhelpful to connect an ill-defined problem (consciousness) with theories that are hardly expressible in terms relevant to cognitive neuroscience (Segalowitz 2009 ; Smith 2009).

[139] Schaffer objects more principally to the assumption of a fundamental level and argues that the distinction of levels or hierarchical structures does not imply acceptance of this assumption (Schaffer 2003).

[140] Moss's critique of mechanistic explanation is partly directed against the contrast with nomological explanation that several proponents make. He underestimates the potential for combining the two forms of explanation, though. Furthermore, the following statement demonstrates that he overlooks the fact that an explanatory mechanism is relevant only for a particular phenomenon: "the presupposition of any functional, let alone mechanistic, analysis is the holistic assumption of a unified entity that acts flexibly and contingently to sustain its own existence" (Moss 2012 166).

# 6  CONCLUDING REMARKS AFTER CONSIDERING THE FOUR METHODOLOGIES

With the explanation of dynamic changes of a cognitive function, we have probably touched upon the most difficult topic both for empirical explanation and for its philosophical analysis. This difficulty already transpired when we acknowledged that it raises questions of identity and difference that belong traditionally to the metaphysical realm. To the extent that these questions touch upon the observation, interpretation and explanation of behavior we can even recognize normative issues at stake: are bird song or infant crying forms of singing, can automatized behavior be a form of moral action?

These remarks implicitly account for both our acceptance and rejection of parts of the argument discussed in chapter I.2, where we considered the role of conceptual analysis in cognitive neuroscience. Even though we did gladly acknowledge the importance of definition as a heuristic and as a form of constraint on scientific research, the neglect by Bennett & Hacker of the pluralism in classifications and definitions, and of the problem of blurred distinctions rendered their view of conceptual analysis inapplicable. Moreover, as a consequence of their assumption that singular definitions on the basis of conceptual analysis and behavioral criteria are possible, the authors strictly demarcated conceptual from empirical work, as was demonstrated by their quote: "[c]onceptual truths delineate the logical space within which facts are located. They determine what makes sense. Consequently facts can neither confirm nor conflict with them" (Bennett and Hacker 2007 129). However, once the assumption of a strictly delineated logical space on the basis of conceptual truths is abandoned, the relation between conceptual analysis and empirical science must be redesigned.

Unsurprising to most of us, cognitive neuroscience is a highly interdisciplinary field where the integration of insights is required. The role for conceptual analysis or definition still needs to be determined, as does the form of its integration with empirical insights. Both were found to be relatively loose in the approach that aims at neural correlates of consciousness. Probably because the concept or definition of consciousness is notoriously problematic in itself, the search for its neural correlates was found to be performed without a preliminary definition of consciousness. In fact, the NCC approach partly aims to circumvent that problem by seeking to employ empirical evidence alongside concept analysis as a way to delineate the phenomenon of consciousness more precisely. As a case in point we referred to the existing analogy between the proposed neural correlate of recurrent processing (Lamme 2006 499)

and the definition of consciousness as reliving or rekindling experience (Dennett 2005).

The other two approaches were more articulate about the requirements for a preliminary definition of the function or phenomenon to be investigated, and the role of such a definition. In the case of Marr's methodology, it was a more precise task analysis that should help to delineate further investigations. The role of further empirical evidence for this computational theory was somewhat ambiguous in his approach, we found. Even though Marr suggested to keep the three levels or theories relatively independent, he himself actually considered mutual constraints between these three levels, if only as constraints on the search space for options of the neighboring level.[141] A particular algorithm theory, for example, would be best served by a particular neural implementation, which subsequent empirical research could try to determine, or vice versa (Marr 1982). This modest integration of insights stemming from different levels or perspectives of research was found to be even more elaborate in the mechanistic explanation approach.

Employing several of the heuristics or methods that were discovered in the other three approaches, mechanistic explanation appeared to be both relatively modest in its ambition – as an explanatory mechanism is primarily relevant to a particular phenomenon – and explicit in its performance and requirements. For example, it requires researchers to delineate or define the phenomenon of their research, while leaving them room for its later redefinition or reconstitution. As we argued, such redefinition or reconstitution can be the useful result of the detection of particular properties or constraints of the phenomenon's explanatory mechanism. Consequently, research leads to a continuous integration of results in the development of an increasingly elaborate mechanistic explanation. More than the other approaches, mechanistic explanation is suitable for such an integration, since it can handle not just simple but also complex phenomena and include dynamical aspects like development and learning.

To a large extent, therefore, mechanistic explanation collects resources and tools to handle and integrate elements that were present in the previously discussed explanatory approaches. Moreover, with its particular interest in the organized dynamics of a mechanism that produces a complex phenomenon, this approach can facilitate the investigation of the subject of this dissertation. Focusing as it does on the complex dynamics of action determination, to which different components

---

[141] In fact, as we noted earlier, Marr used all three levels to constrain each other. Kosslyn accordingly arranged the levels in a triangular fashion, ascribing equal weight to all three (Kosslyn and Maljkovic 1990).

contribute, both in terms of parts and of operations, the resources of mechanistic explanation help in formulating questions and organizing results. As philosophical – conceptual – analysis will teach us that the determination of action involves both explicit decision making processes as well as tacit automatisms and habits, we can expect a variety of components involved in constraining or sculpting the space of actions depending upon the relevant mechanism. Moreover, the explanatory mechanism also allows us to elucidate the transformative impact of learning, of automatization and habit formation on this space of action by referring to its components and their organization and interactions.

Given the importance of the modification of an explanatory mechanism under the influence of development and learning, the next part will be devoted to that topic. We will observe different explanations for the modification of a mechanism responsible for a particular function. Such a modification can occur, so we will learn, under the influence of information or concepts that somehow become integrated in the responsible mechanism; information that may be represented in the environment, but also employed by an agent himself as a form of self-regulation. Part II is meant to support the hypothesis that not just simple but also complex functions are suitable for automatization and habituation, which will eventually affect their properties – like the automatization of singing in experts having an impact on its speed, efficiency, control, flexibility, and so on. Discussion of this phenomenon, so prevalent in dynamic systems, will also bring back the questions that we raised at the beginning: must we adjust the definition of a function once we discover that it can be performed under such different conditions and with correspondingly different properties? Does it make sense to talk about 'dual-processes' that allegedly underpin many forms of human cognition and action and to consider these as two distinct mechanisms? Must we then also consider redefinition or reclassification of the phenomena these mechanisms produce? Or should we consider these processes as being produced by a single explanatory mechanism under different conditions and with different outcomes?[142] Again, these questions are not just empirical, but also philosophical and even normative - questions that will not leave us in the rest of this dissertation.

---

[142] Though this is not the place to discuss this, the relevance of neural dissociation research is also at stake here. Is the finding that two distinct processes or mechanisms are involved in different performances of a particular function enough reason to split the function into two distinct functions? Or are such empirical results completely irrelevant to our conception of that function? To the extent that dissociation research relies upon specific assumptions of modularity, its logic is rejected by (Orden, Pennington et al. 2001). However, it is argued that double dissociation research depends upon less strict assumptions and that its evidence can be used as additional support for a particular cognitive theory over others (Davies 2010).

# Part II
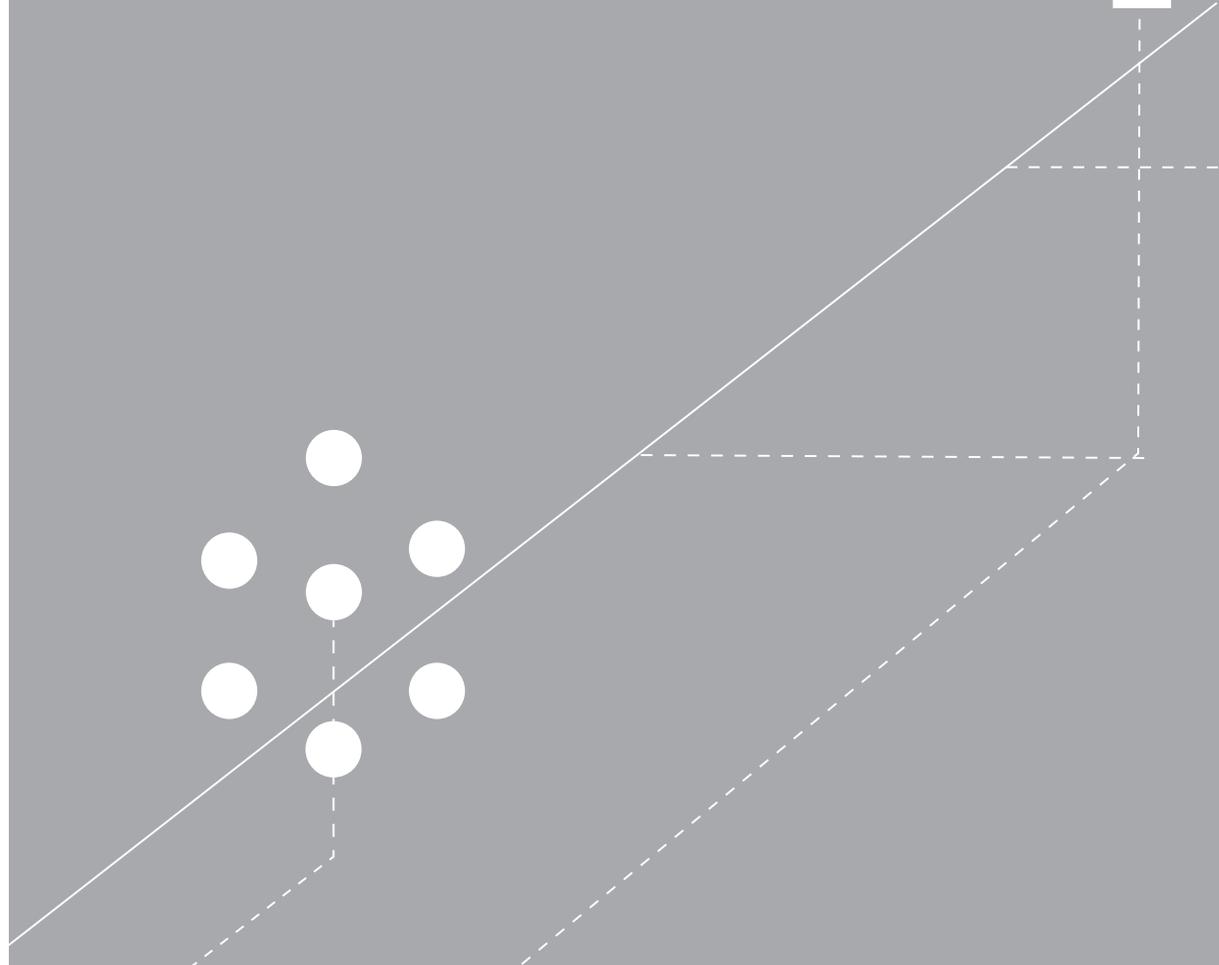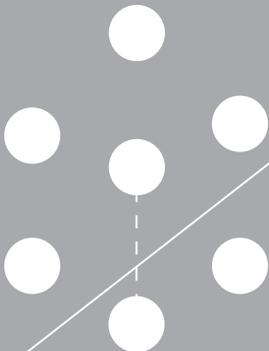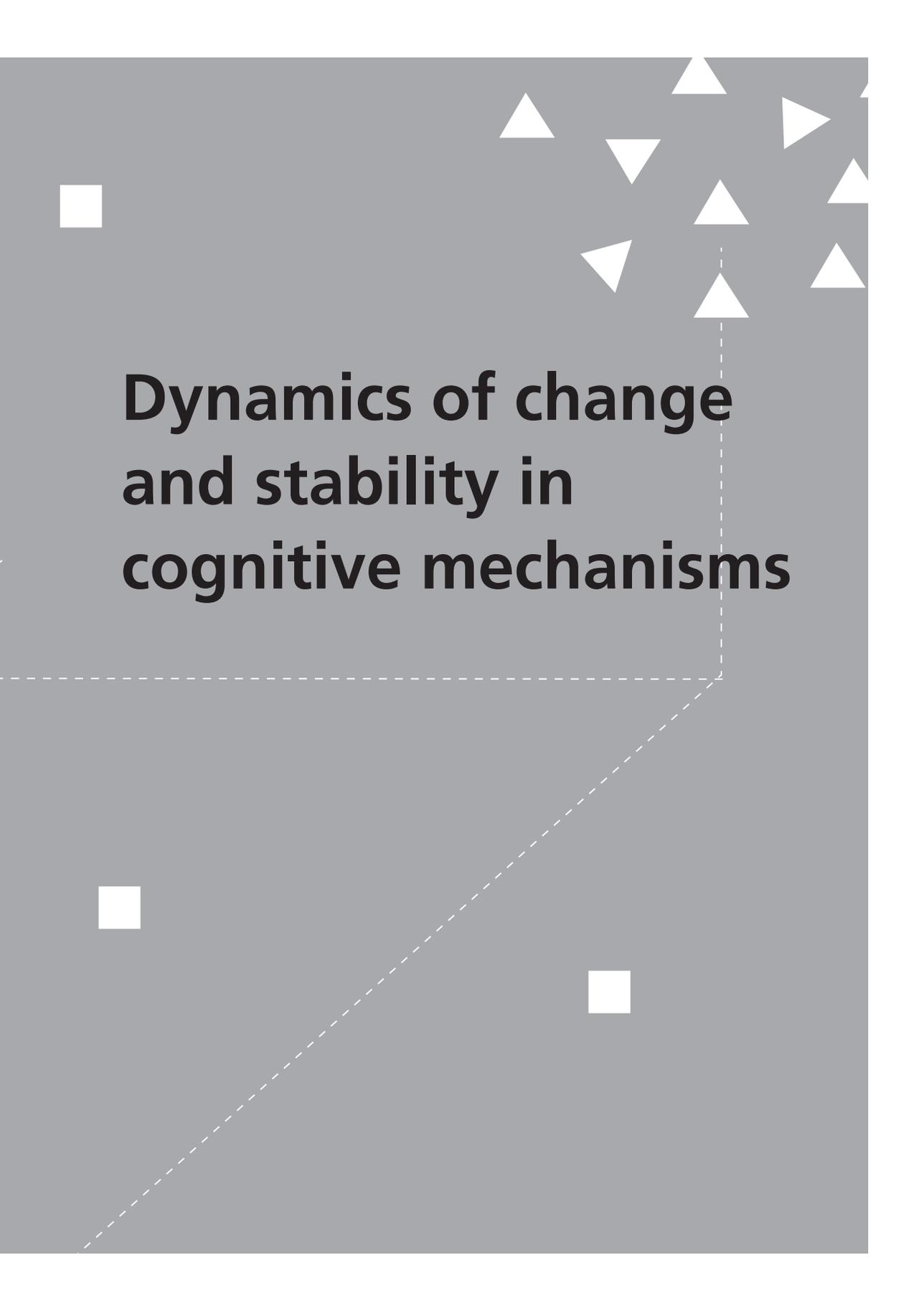
# Dynamics of change and stability in cognitive mechanisms

# 1  INTRODUCTION: FROM DYNAMICS TO STABILITY AND BACK AGAIN*

Hearing the babbling sounds of a baby can be as enjoyable to its parents as a song recital by an expert vocalist. Nonetheless, the babbling is a far cry from the expert singing in many respects, even though every singer will have started as a babbling baby and remains capable of such babbling. In a sense, then, increasing expertise in singing amounts to adding different modes of vocal expression to one's repertory and to enhancing one's control: controlling each mode of vocal expression separately and controlling which mode of vocal expression is to be used at a particular moment. The development along these two dimensions could at first sight appear to be paradoxical: if one enforces upon an agent an increasing number of factors to determine when to perform a particular task, one would normally expect this agent to lose control in comparison to his earlier performance of that – simpler – task and therefore show a deteriorated performance. This apparent paradox notwithstanding, development and learning appear to enable both the enrichment of a particular capability and the stabilization of the performance of that capability, which includes different modes of performing it. This part will focus on the aspects of such a process of development or learning that have to do with the stabilizing and expansion of a particular capability like vocal communication, preparing for our discussion in Part III of the diversity in our modes of performing intentional action. Such a process is complex, as it involves gaining expertise with certain elements of that function, increasing ability to combine different elements of the function with each other and eventually to have such ease with it that the function can be performed alongside other functions. Let us consider the singing example again to clarify this, describing how it becomes increasingly complex and integrated with other functions.

To begin with, babbling lacks language while singing usually combines musical with linguistic elements – sometimes even in a language other than the singer's mother tongue. Apart from these linguistic elements, the musical elements of singing are generally not distinguishable in babbling sounds: rhythm is lacking, dynamical structure is usually very simple and the tonal spectrum is rather narrow.[143] An expert singer, in contrast, has mastered a wide range of distinct rhythms, can apply large dynamical differences to the music and control his voice such that he has added octaves to his vocal range compared with novice singers and babies. Moreover, depending on the intended expression, an

---

* On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively. Fig. II is particularly relevant as a representation of the main contents of section II.2.

[143] As for melody: research with French and German newborns demonstrated that these had already memorized the prosody or melodic contour specific to their mother's language and reproduced this in their own whining (Mampe, Friederici et al. 2009).

expert singer can control not just the pitch but also the timbre and projection of his voice, adding another dimension to his music. Being able to combine those musical elements and linguistic elements at wish, an individual singer can impersonate in the course of a single recital a seductive Don Giovanni, a meditative Saint François, an amorous Tristan, a raging Pizzaro, a joyous Porgy, or a babbling (indeed, babbling) Papageno (or female parts for female singers, obviously). Finally, an obvious difference is of course that singing happens mostly in an ensemble with an accompanist or other singers, while babbling babies usually perform individually. Corresponding with this difference is the reproducibility which is desirable for expert song, enabling a musician to improve his performance individually, to adjust his interpretation of a song or improvisation at wish and in detail, and to play together - often aided by a score of the music, from which he can sing at prima vista. Given these differences between baby babbling and expert song, it is remarkable that the two phenomena are nonetheless related, that the former is a necessary precursor to the latter and does remain on the singer's repertory, with the two standing in a developmental relation to each other. Our discussions in the previous part have provided us with some useful tools to explain and understand this curious fact.

The ingredients of mechanistic explanation, so we argued, are such that we can elucidate the dynamic modifications of a mechanism responsible for a cognitive or behavioral phenomenon. Given a mechanism's complex organization of components parts and operations and the mechanism's interactions with its environment, there are ample resources to account for such dynamics. Indeed, section I.5.6 presented four possible modifications of a mechanism: 1) the recruitment of a novel component part in the mechanism responsible for a phenomenon like singing, as when babbling gets combined with linguistic elements; 2) a dynamic change in a component operation of the mechanism, as when a singer is able to correct his pitch very fast; 3) a change in a mechanism's organization, as when singing has become so automatized that an opera singer can shift his attention to his stage performance; 4) a modification and expansion of the kinds of interactions with the environment, as when an expert singer is capable of colouring his voice such that it drowns out different types of accompaniment and fills music halls with different acoustics. These modification types will usually co-occur, as it is often impossible for a modification type to obtain separately. A relevant example is the fact that usually, a novel form of environmental interaction with another singer is only possible when some singing abilities have previously stabilized.

With these ingredients put in place, the question may arise whether a phenomenon that is produced by such a modifiable mechanism is not in constant flux. After all, the activities of a mechanism and the associated environmental and internal interactions

continuously affect its components and their organization, so how can we still expect it to perform stably? Indeed, these effects on the mechanism will not always be identical or consistent with each other, as every practising singer knows very well, hitting false notes all too often. Still, this modifiability does not preclude a singer from obtaining stable results in his singing after a period of practice. After repeating a vocal coloratura many times and in segments at first, the singer will then be able to reproduce it reliably. This stability of his performance enables him subsequently to speed up this part, to vary its dynamics, to transpose it a semitone when his duet partner forces him to do so and even to draw the specific facial expressions required by a stage director. Instead of having to struggle when singing this part from note to note, the singer now feels as if the coloratura has become a single unit,  allowing for as much freedom with that coloratura as a novice singer may enjoy when singing just a single and easy note.

This phenomenon of increasing expertise with a certain cognitive or behavioral function or an aspect of such a function is familiar to all animals. Immediately after birth the functions to be developed and stabilized have to do with feeding and moving, followed by on-going processes of improvement of those functions and varieties thereof and of learning completely new functions. The set of functions that an animal can perform is constantly changing and generally expanding, with subsets becoming increasingly stable. Now that Part I has provided us with the tools to explain a particular function and its underlying dynamics, Part II seeks to analyze and explain how increasing expertise affects a cognitive function. For this, we will take up three accounts that all present different explanatory mechanisms underlying such expertise or skill learning. From our perspective, however, these accounts agree in that expertise has an effect on the responsible mechanism, an effect which we will consider in terms of 'kludge formation'.

To do this, we will first devote a chapter of this part to the definition of a 'kludge'. Since our intention is to argue that learning and development results in novel cognitive or behavioral response patterns on the basis of the dynamic changes in a responsible mechanism – characterized by a hierarchical and modular structure - that correspond to the formation of a novel kludge in it, this concept will play an important role here. We will then devote three chapters to cognitive phenomena that represent such processes. One of these will concern the development in children of domain specific or task specific brain circuits that can subsequently function as kludges, a process that has been referred to as 'modularization' (Karmiloff-Smith 1992). A second chapter will discuss so-called 'dual-process' theories, referring to two different and allegedly independent processes - or for some authors: systems - in the human brain, operating on different informational domains and according to different rules, which allegedly

leads to inconsistent behaviors (cf. Chaiken and Trope 1999 ; Evans 2008 ; Lieberman 2007 ; Smith and DeCoster 2000). Chapter II.4, finally, will discuss the theory that humans are capable of forming cognitive kludges not only by combining different capacities, but also by including external components like tools or symbol use in such kludges (Barsalou 1999c; Clark 1997). Discussing these examples will pave the way for Part III, which will be devoted to the complex process that is involved in action determination, including the use of language. This process is to a large extent carried out by a set of kludges, varying in their origin and functional properties, which contribute in multiple ways to the process of 'sculpting the space of actions'. It is to the definition and characterization of a 'kludge' that we now turn.

## 1.1 Kludges: mechanism adjustments and expansions

Above, we noted that due to development and experience, an expert singer has become capable of performing well and stably at more levels of specificity and complexity than a novice singer, also allowing him to expand his performance both in vocal terms as in other terms like playing a stage character or keeping track of an accompanist. This capability is dependent upon a process that yields a mechanism comprising ever more components of a type that we will from now on call 'kludges'. As we noted in Part I, a mechanism generally has a hierarchical – or rather: heterarchical[144] - modular structure: it is characterized by a nested structure of components with relatively specific functions, each contributing to the mechanism's performance as a whole. Such a hierarchical and modular structure can undergo several types of modifications, as we mentioned. Whether through normal development, via specific learning episodes or via common accumulation of practice, the modification of the relevant mechanism's structure will usually involve the emergence of a 'kludge'.[145] In such a case we may also refer to the implicated mechanism as being 'kludged' and the function it produces –behavioral or cognitive- as being a 'kludged function'. Even though a kludge bears some resemblance to what is commonly referred to as a 'module', we prefer the former term in order to avoid some of the undesirable associations with the latter term – about which more below. What then are the defining characteristics of a kludge?

---

[144] That is: a heterarchy with modifiable control relations – see footnote 96 in part I.

[145] The term 'kludge' or 'kluge' is common within the engineering domain, where it refers to what Marcus calls 'a haphazard' construction (Marcus 2008), or a 'cobbled together' solution in Clark's terms (Clark 1987). Its origin and precise spelling is unclear. Marcus refers to the German word 'klug' for clever as a potential source (Marcus 2008), reason why he spells it without a 'd'. Although the German association is thought provoking, spelling it with a 'd' additionally maintains the association with the English word 'clutch', referring to a mechanism that draws in an additional component for fulfilling a function.

To begin with, a kludge is a component of a mechanism, which must be characterized functionally. That is to say, when a kludge emerges we can observe this in changes in the performance of a particular cognitive or behavioral function or component function, related to a specific domain. As we will see below, usually the performance of such a function – which here implies also: component function, as it often applies to a specific function like keeping tone or singing a coloratura – happens with greater speed, stability and flexibility than it used to happen earlier by the subject. This was the case with much of the singing components discussed above, whether it was reading notes, singing coloraturas or other function components. Depending on the specific function, the functional properties that change due to its being performed by a kludged mechanism obviously will differ. Sometimes, for example, it may not be visible in changes in the kludged function, but we could perhaps witness from changes in unrelated functions that cognitive resources for the latter have become available due to the reduction in resources required for the former, now kludged function. Obviously, we may assume that changing properties in the kludged function are to some extent determined by the algorithmic processes or the neural implementation that are associated with this kludge formation. Let us turn to the second important aspect of what makes a kludge: its algorithmic processes.

The emergence of a kludge related to a specific function usually corresponds with a change in the algorithmic processes involved. Nonetheless, the second characteristic of a kludge is that we cannot directly derive from its functional – cognitive or behavioral – properties an 'algorithmic' theory in Marr's sense, as discussed in section I.3.3.1. (Marr 1982). Indeed, the mere fact that development or learning affects the processes involved in a function like singing demonstrates that such a function can have a multiplicity of processes subserving it. One and the same functional result can be obtained via more than just a single process, involving different types of information processing, or representation manipulations, or dynamical processes. In singing, for example, practicing a particular difficult vocal phrase can involve its segmentation in smaller parts and gradually connecting these, or developing a mental image of the phrase and thus facilitating vocal muscular movements, or simplifying the phrase and gradually reinserting the difficult parts, or imitating a vocal expert's examples which will often be sung with increasing speed. Clearly, all of the algorithmic processes involved are associated with other cognitive or behavioral functions, some of which are rather complex. Perhaps score reading is involved, which can help the segmentation of the phrase in smaller parts. Or careful listening is important, when imitation is involved. Not only can these algorithmic processes differ between subjects, these processes will generally also change as a function of the kludge formation. Indeed, in many cases the

formation of a kludge appears to result in a change of precisely this component, of the algorithmic processes involved in a particular function. This is the case, for example, when a response to a particular situation no longer requires complex cognitive and perhaps conscious processing but simply is given immediately upon perceiving a task specific stimulus due to its being habituated (Graybiel 2008). In many other cases and several of the examples to be discussed below, however, kludge formation is associated with specific changes in the representations that are relevant for the task at hand, like the musical representations in singing.

Now that we have argued that it is not possible to derive a specific algorithmic theory from the fact that a function and its underlying mechanism have become kludged, it is obvious that this also holds for the specific neural implementation of a kludge: its third characteristic is that there may be more than just a single option available for its neural implementation. As we learned in Part I, any investigation of the neural implementation of a function relies on the preliminary definition of its functional properties. Marr already pointed out that it may even be possible that one and the same function – characterized in his approach with a computational or task theory (Marr 1982) – allows different kinds of implementation. This is even more so with a function that is involved in such kludge formation, as this process implies the modification of the mechanism responsible for the function, even though the associated modifications may differ from case to case. Indeed, it may be the case that while most subjects will establish a kludge when learning to perform a particular function, like finding a note or singing a coloratura, the specific neural process that corresponds to its formation may vary between stages of the kludge formation process, or vary between subjects. This has partly to do with the fact that the performance of a particular function can rely on different cognitive or behavioral processes and consequently the underlying neural processes involved. Regarding those neural processes, in section I.5.6 we discussed that there are several ways in which an underlying explanatory mechanism can modify due to learning and development. In terms of neural implementations, we mentioned the option of structural changes of the responsible mechanism, or the option of connectivity changes within a mechanism of which the components and operations remain largely the same.

A fourth kludge characteristic pertains to the fact that there will be quite some variation between subjects or even within an individual subject during the intermediate stages of learning or development, even when a particular kludge formation might eventually lead to rather similar functional, algorithmic, and neural implementation properties. Particularly in experienced subjects, we can even assume that more than one kludge can be activated or employed for the performance of a particular function.

Indeed, one of the consequences of their being expert singers is that they are capable of engaging different kludges for the performance of a particular function, depending on some other, relevant functional properties. Perhaps in a situation where a singer is tired, his singing will primarily rely upon a particular kludge that allows him to perform his score well, even though this kludge allows less adjustment of tone to the sound of other performers, it leaving the singer less responsive to auditory input.

A further defining characteristic of a kludge is that it often consists of components – functional, algorithmic, neural – that are already present but are then 'cobbled together', a feature from which a kludge in fact derives its name (Clark 1987, cf. Marcus, 2008).[146] Thus, a kludge will often have properties suggesting that its emergence is primarily due to development or learning and not genetically determined or innate. Although a kludge generally consists of a mechanism modification in which development or learning were involved, it is usually the result of not just one but a combination of different constitutive forces. It may involve a modification of a speech component mechanism after a specific learning period, or it consists of a modified music score reading mechanism that involves the adaptation – or recycling (Dehaene 2005) – of among other things the Fusiform Face Area. As a result, it is often hard to determine the developmental, experiential, environmental and other factors contributing to such an emerging kludge.

A sixth characteristic of a kludge is that after its emergence it can itself become involved in subsequent developmental or experiential trajectories. Notwithstanding the fact that the emergence of a kludge involves the modification of a responsible mechanism, this newly emerged mechanism or mechanism component itself will probably play a role in subsequent developments. Indeed, although a kludge generally emerges from strongly associating components that were already in place, it may itself in turn become similarly involved in another kludge formation process. As a result of this, one can observe that some kludges have become so deeply entrenched in other mechanisms with specific functions, that its disturbance would have wide-ranging consequences and not be limited to the kludge's specific functional characteristics.[147]

---

[146] Concurring with this characteristic are several accounts of the development of cognitive and behavioral functions that involve extensive neural re-use in the brain (Anderson 2010), or the exploitation of previously established 'neurofunctional architecture' for new functions (Gallese 2008), or the 'recycling' of older brain circuitry for cultural inventions like reading (Dehaene 2005).

[147] Wimsatt and Rasmussen both point out that it may be useful to distinguish between the degree of 'generative entrenchment' of particular properties as a measure for their being involved in other, later developments, as it allows us to distinguish between properties with a more recent or an older evolutionary or developmental history (Rasmussen 1987 ; Wimsatt 1986 ; Wimsatt 2001). Karmiloff-Smith argues from a different perspective – about which more below in chapter II.2 – that functions that 'modularize' at an early age will have a greater impact when impaired than others, making double dissociations for modular components highly improbable (Karmiloff-Smith, Scerif et al. 2003). Her reservations against the common definition of 'modules' is one of the reasons we prefer the notion of 'kludge' in our argument.

It is not surprising that at times researchers harbor the unwarranted assumption that the components of a particular kludge must themselves have a 'natural' origin or be innate, instead of these also being the product of a previous idiosyncratic process of kludge formation.[148]

A seventh and final characteristic of kludges that we need to mention is the involvement of external, environmental information in the process of their emergence and in their functioning. As a kludge often emerges after a period of specific experience or learning in which a subject engages in a particular manner with his environment, its functional properties –mentioned above as its first defining characteristic – often include environmental information. Such an inclusion can range from a kludge's activation by a particular environmental stimulus to the inclusion of the properties of a particular tool in one's body scheme.[149] Even though some kludges emerge at such a young age in subjects and involve the inclusion of such elementary environmental information that they have sometimes been interpreted as innate mechanisms, their properties cast some doubt on the strict distinction between a kludge's being innate or acquired.[150]

Now that we have spelled out these kludge characteristics, let us turn to the first of three examples in which kludge formation can be associated with observable changes in cognitive and behavioral functions. Since we will not find reference to 'kludge' or to kludge formation in this case, but rather to the concepts module and 'modularization' (Karmiloff-Smith 1992), we will start the next chapter with a short discussion of these concepts.

---

[148] An interesting example is the debate whether mirror neurons, involved in the mirror neuron systems that are taken to be responsible for many intersubjective processes, are innate or the result of a learning trajectory. It appears that consensus is increasingly in favor of an experiential basis of mirror properties (Catmur, Mars et al. 2011 ; Keysers and Perrett 2004).

[149] Iriki among others suggests that associated with this process of incorporating or assimilating of tool properties in one's body scheme, an analogous process of 'objectification' of one's body takes place, facilitating imitation and mutual learning (Iriki 2006).

[150] Wimsatt in particular has warned against this strict distinction between the innate or acquired nature of – what he calls – 'generative entrenchments'. Elaborating on Mayr's notion of closed and open behavioral programs (Mayr 1974), Wimsatt refers to the example of imprinting, which is a tightly constrained mechanism that nonetheless includes environmental information in its eventual emergence –potentially including information associated with ethologist Lorentz in his goose chicks' imprinting mechanisms (Wimsatt 1986).

## 2  MODULARIZATION AS A PROCESS CORRESPONDING TO LEARNING AND COGNITIVE DEVELOPMENT*

The reader may have been asking himself why we did not simply adopt the common notion of 'module', which largely overlaps our concept 'kludge'. Indeed, the fact that we subscribe to the notion that explanatory mechanisms for cognitive functions – and many others as well – generally have a hierarchical and modular structure would support the adoption of the former. However, there are a number of reasons not to do so.

From its seminal discussion by Fodor in his 1983 publication of 'The Modularity of Mind' (Fodor 1983), the notion of module has received quite different interpretations. Fodor defended the presence of modular systems particularly involved in perception and language processing –at work in the periphery instead of the central systems of the brain, in his words. He suggested that such systems in the brain respond to some or all of the following conditions: the input system is domain specific, its operation is mandatory while allowing limited top-down access, it is fast, its operations do not rely on other sources of information, its output is rather shallow, it is performed by a 'fixed neural architecture' and recognizable by characteristic lesion patterns and finally it displays a characteristic developmental trajectory. The relevance of this notion is made clear with the following remark in the "Caveats and Conclusions" to his essay: "the limits of modularity are also likely to be the limits of what we are going to be able to understand about the mind" (Fodor 1983 126).[151]

Notwithstanding the relevance of the notion, there has not been general agreement regarding  its definition or its empirical plausibility. A first obstacle has been that many authors interpreted Fodor in the sense that the complete list of conditions for modularity must be applicable for a particular system to be modular[152] – even though Fodor did not require this (Coltheart 1999). Consequently, authors could dispute the applicability of the notion in a specific case by merely referring to a single condition's not being fulfilled in that case.

A second obstacle is that many authors associated the notion of 'module' with an additional condition, namely its innateness. This was not mentioned as a separate

---

* On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively. Fig. II is particularly relevant as a representation of the main contents of section II.2.

[151] This caveat concurs with the methodological emphasis that Simon and Marr put on a 'near-decomposable' (Simon 1962) or modular (Marr 1982) structure for a system to be explained – both authors being aware that it is unclear whether this requirement refers to an epistemological or a metaphysical demand, or perhaps both.

[152] As for instance happens in the relatively positive review article on the use of modularity and its future by (Thomas and Karmiloff-Smith 1998 246).

condition by Fodor, who referred only to the possibility of an innate specification of the information that a module is capable of handling (Fodor 1983). Nonetheless, this association with innateness led to the reproach that Fodor defended an 'anti-constructivist nativism', which would correspond with his underestimating the plasticity of the human brain and its influence on the formation of modules (Karmiloff-Smith 1994).[153]

A third obstacle has been Fodor's denial that modules can be assembled: "in the sense of having been put together from some stock of more elementary subprocesses" (Fodor 1983 37). This would limit the number and kind of modules severely and restrict the applicability of the term largely. Arguing against this contention, Coltheart shows that Fodor's own descriptions of particular language processing and visual information processing modules in fact contain references to modular component processes like a phonetic processor or a 'form-concept dictionary' (Coltheart 1999).

In more recent years, the notion has become not only widely used but also interpreted ever more loosely. Trying to capture the debate, a recent review suggests interpreting the different notions of modularity along five different dimensions used to determine a case of modularity: its physical structure, its cognitive functions, its information processing or computation, the information it employs and finally its development. The author's attempt to find consensus in the debate about the most relevant dimensions – computational, informational and physical – seems implausible in light of the divergence he noted earlier (Seok 2006).

A more recent attempt at articulating modularity, together with an explication of its prominence in systems of various kinds, seems more promising to us. The notion of modularity has regained interest due to large scale brain imaging and computational studies that support the prominence and effectiveness of hierarchical and modular structures for the explanation of brain activation patterns and corresponding cognitive functions. Specifically with regard to the networks that underlie these activities and the topological structures of these networks, it appears that these structures are modular at several levels. That is to say: "many systems have the fractal property of hierarchical modularity, multi-scale modularity or "Russian-doll" modularity (Meunier, Lambiotte

---

[153] Generally, authors who defend the position that a module may be the result of some ontogenetic developmental or learning trajectory, have had difficulty with the notion of modularity. Reviewing most of the recent, relevant literature on the notion, Barrett & Kurzban conclude that it has been the "equation of modular with "fixed," "innate," and "static"" that has led to much confusion (Barrett and Kurzban 2006 642). Marcus also argues that both the notion of modularity and the notion of the innateness of cognitive functions are implausible and unproductive but are nonetheless often combined by authors. In contrast he relies on insights from recent approaches in the study of evolution & development that suggest that the brain is determined by both 'prewiring' and 'rewiring': "innateness is about the extent to which the brain is prewired, plasticity about the extent to which it can be rewired" (Marcus 2009 151)

et al. 2010 2). Reason for this prevalence is that such systems are rapidly and robustly assembled, according to the authors – demonstrating that their approach has implicitly removed the three obstacles that we mentioned. They note, however, that as much as modules can be assembled over time, ageing is also associated with changes in relevant hierarchical and modular structures and corresponding functional changes (Meunier, Lambiotte et al. 2010). With that they emphasize the dynamic nature of modularity, which will be part of our argument below. We will start with evidence from development cognitive neuroscience for the process of modularization and its involvement in the development of cognitive functions.

## 2.1  Neuroconstructivism and the relevance of modularization

Referring to 'Fodor's anti-constructivist nativism' in her seminal developmental cognitive neuroscience publication "Beyond Modularity: A developmental perspective on cognitive science" (Karmiloff-Smith 1992), the author positions herself against a nativist view[154] that involves rejecting the possibility for a brain to function and learn at all if it was not born with some prespecified contents and processes already available in domain specific modules.[155] In contrast she and other authors defend a position which is aptly called 'constructivism' (Quartz 1999) – or 'neuroconstructivism'. The latter position considers development and learning crucial processes that contribute not just to increasingly complex cognition but also to the 'construction' of an increasing complexity of the brain's networks that includes their gaining in modular structures (Karmiloff-Smith 2009 ; Mareschal, Johnson et al. 2007 ; Westermann, Mareschal et al. 2007).

---

[154] Nativist positions with respect to modularity gained strength with contributions from evolutionary psychology as it presented evolutionary explanations for the brain as an evolved container of distinct expert systems (Cosmides and Tooby 1997) – leaving an all too modest role to individual development and thus inviting criticism that insists on modularization and a constructivist account of development (Wheeler and Clark 2008). Evolutionary psychology has been criticized among other reasons for its insistence on innate modularity – peripheral, massive or otherwise (Buller and Hardcastle 2000 ; Prinz 2006). Differentiating developmental, functional and mental modules while assigning a role to development in evolutionary psychology, Griffiths contends that the modules investigated in neuropsychology and in evolutionary psychology may differ (Griffiths 2007).

[155] To be precise, her neuroconstructivist position challenges both "Fodor's anti-constructivist nativism and Piaget's anti-nativist constructivism" (Karmiloff-Smith 1994 693) – leaving some role for innate constraints that partly determine further developments. This position is similar to the 'open programs' proposed earlier by Mayr (Mayr 1974) and further elaborated by Wimsatt (Wimsatt 1986). However, as mentioned above, Fodor did not require all conditions to be met for a case of modularity: "it is reasonably easy to think of psychological processes that are fast but not encapsulated, or involuntary but not innate, and so forth. The present contention, in any event, is relatively modest" (Fodor 1983 137, note 35). Nonetheless, Karmiloff-Smith overlooks this when she writes: "It is the co-occurrence of all these properties that constitutes, for Fodor, a module" (Karmiloff-Smith 2006b 10). The identification of Fodor's position with 'nativism' is therefore exaggerated.

Observations of children learning skills in language, physics, mathematics and psychological reasoning are discussed in Karmiloff's 1992 plea for a 'developmental perspective on cognitive science'. Other skills, like the motor skills of playing the piano and solving Rubik's Cube, are treated from a similar perspective, even when adults rather than children are the learners here (Karmiloff-Smith 1992). That same developmental perspective was later applied to neurological disorders like the Williams syndrome or Specific Language Impairment, on the basis that such a perspective can better account for the widespread symptoms of these disorders than perspectives which assume that either a brain operates without modular structures at all or instead that a brain is innate with a modular structure that is similar for newborns and adults alike (Campos and Sotillo 2008 ; Karmiloff-Smith 1998 ; Karmiloff-Smith 2009 ; Karmiloff-Smith 2011 ; Karmiloff-Smith, Scerif et al. 2003 ; Mareschal, Johnson et al. 2007). As we are not specifically interested in disorders nor in infant development, we will limit our discussion here to those aspects of development and learning that appear to hold for adult learning, too. Then we will ascertain whether the seven kludge characteristics listed in the previous chapter, apply to it.

Observations show that development and learning consist of two distinct processes that to some extent even seem at odds with each other: proceduralization and explicitation.

Learning usually starts with the proceduralization of a particular skill, amounting to: "rendering behavior more automatic and less accessible" (Karmiloff-Smith 1992 17). This is complemented by its explicitation process, involving representations of the skill's domain. These representations contain increasingly "explicitly defined" components, which offer the child new opportunities for adjusting its performance, for example because: "the potential relationships between procedural components can then be marked and represented internally" (Karmiloff-Smith 1992 22).[156] According to the neuroconstructivist account, development and learning depend on 'Representational Redescription' of representations of the relevant domain of the skill, with the initial representations being implicit and the subsequent three stages involving increasingly explicit representations.[157]

The term 'modularization' refers to elements of these two processes insofar as they imply that: "input and output processing becomes less influenced by other

---

[156] This focus on representation has been maintained in a more recent neuroconstructivist account of development, distinguishing itself from a dynamic systems approach: "we consider representations as the central construct of cognitive development" (Mareschal, Johnson et al. 2007 89).

[157] This idea of a series of representations is akin to Marr's idea that visual information processing involves the processing of a series of successive representations of the retinal image until a 3 D representation has been formed (Marr 1982).

processes in the brain. This causes knowledge to become more encapsulated and less accessible to other systems" (Karmiloff-Smith 1992 15). Although modularization applies particularly to the proceduralization component of learning, there is a direct connection to the representational changes involved in such learning and the representations subsequently becoming available for explicitation, for correction, for their use in other skills, and other uses (Clark and Karmiloff-Smith 1993).[158] Indeed, this connection has been rendered in a more recent neuroconstructivist monograph as: "representations trapped within modules underwent an intrinsic process of abstraction even after behavioral mastery, which would eventually 'offer up' the knowledge within the module to other cognitive processes" (Mareschal, Johnson et al. 2007 213).[159]

These short descriptions of the processes involved in development and learning according to a neuroconstructivist approach may suffice for now, as our main goal is to consider whether this account concurs with our argument that the formation of kludges play an important role in development and learning generally.[160] Let us therefore consider the seven kludge characteristics mentioned in section II.1.1 and compare these with evidence stemming from research according to this approach.

## 2.2  Modularization and the seven kludge characteristics

The first issue pertaining to a kludge was that, although it is part of an explanatory mechanism, it must be characterized primarily functionally – its emergence should be visible in the cognitive or behavioral responses of an individual in a particular domain. The neuroconstructivist approach to cognitive development defended in (Karmiloff-

---

[158] The correspondence between increasingly separate and isolated processes with increasing need of specific processes aiming at the interaction of such separate processes is in a review called the combination of 'dissociation' and 'integration' in cognitive development – itself again associated with neural developments (Johnson and Munakata 2005).

[159] The authors identify the prefrontal cortex as likely involved in this process of interaction and mediation between specializing modules (Mareschal, Johnson et al. 2007). This is indeed a plausible candidate for such a role, given the fact that prefrontal cortex is highly connected to many long-distance neural areas  involved in the processing of rules and several linguistic elements, but is not engaged by 'routine, automatic and overlearned behavior' (Fuster 2001).

[160]  Comparable in these respects with the neuroconstructivist account of development and learning is the account of implicit learning developed by Cleeremans and colleagues. Its main argument is that implicit learning involves changes in the relevant representations that only gradually become accessible for conscious and verbal processing, with cognitive and behavioral improvements obtaining already at earlier stages. In this account, too, modularity is not prespecified at birth and a static phenomenon, but rather dynamic and emerging phenomenon, associated with learning (Cleeremans 1993 ; Cleeremans 1997 ; Cleeremans and Jiménez 2002). Indeed, reference is made to 'functional modularity', leaving undetermined whether a neural form of modularity is associated with this functional characterization of learning a cognitive or behavioral task. Similar to the neuroconstructivist account, this implicit learning account also assumes that learners possess and control an increasing number of different representations of a particular task, yielding them correspondingly more options for performing it either automatized but hardly conscious or conversely conscious but not as fast (Cleeremans 1997).

Smith 1992) is indeed primarily based upon observations of such functional changes, involving the observed processes of proceduralization and explicitation mentioned above. Although aiming primarily to account for the specific details of these observed processes, the approach also seeks to provide explanations in terms of the changing representations involved, and to suggest possible underlying neural processes. An example of such specific observable results is the fact that these parallel processes do not lead to steady improvements regarding both, as an increase in behavioral errors occurs during phase 2, due to the subject's disregarding some external information in comparison to earlier and later phases (Karmiloff-Smith 1992).[161]

The current modularization account concurs with the second kludge characteristic, referring to the impossibility of deriving a particular algorithmic theory from the functional characterization of a kludge or to the possibility of there being more than one algorithmic theory that can account for a specific case of kludge formation. Nonetheless, this account does assign a crucial role to information processing and representations, as was noted in the previous section where we discussed the process of explicitation, it being complementary to the process of proceduralization.[162] Both these processes allegedly rely on several phases of Representational Redescription. The first implicit and three consecutive explicit phases are characterized by different representations of the same cognitive or behavioral task, varying with regard to foregrounded or abstracted task aspects, being increasingly context-free, and so on. The redescriptions can help to explain the differences between phases in cognitive and behavioral responses like the degree of awareness or verbal rendering of these responses. Finally, the account contends that this Representational Redescription process occurs for each task domain separately and perhaps with different tempo and timing (Karmiloff-Smith 1992),[163] while elsewhere it is elaborated how this

---

[161] An explanation for behavioral inconsistencies – or temporary setbacks - associated with a gradual improvement of a task's representation by a child is provided in terms of a bifurcation stage, signaling a non-linear process leading to qualitatively new knowledge in (Raijmakers 2007). A correlation between neural dynamics and inconsistencies in rule learning has been found that is in accordance with this computational result (Durstewitz, Vittoz et al. 2010).

[162] The complementarity of these processes transpires from the observation that development can be considered as: "a progressive increase in the complexity of representations, with the consequence that new competences can develop based on earlier, simpler ones" (Sirois, Spratling et al. 2008 322). The authors distinguish their take on representations – focusing on their increasing complexity – as being different from the earlier neuroconstructivist account that is central in our present discussion. This account is more focused on the increasing abstractness of the novel and differently formatted representations, which should help explain their involvement in different cognitive and behavioral procedures (Karmiloff-Smith 1992).

[163] There may be spill-over effects between domains, as was found in bilingual children who demonstrate more flexibility in drawing non-existing objects than their monolingual peers do, which is thought to rely on the former children's' expertise with handling different linguistic representations of identical objects (Adi-Japha, Berberich-Artzi et al. 2010).

Representational Redescription process provides a plausible model of information processing that can help explain some recurrent results of development and learning, including temporary regress in proficiency of a certain task, initial difficulty for an expert to explain his performance, and integration of expertise with other tasks (Clark and Karmiloff-Smith 1993).[164] It can be concluded, therefore, that there may be more than one algorithmic theory involved in a particular case of kludge formation or in a performance that relies on such a kludge. Nonetheless, even though it is impossible to derive from a particular kludged performance which algorithmic theory or representation is involved, this does not rule out the possibility of influencing this kludge formation process or performance by choosing a particular representational format of a given task. Obviously, such a format choice will have wider consequences as it may impact upon associations with other tasks or processes, as when a particular music representation format is chosen. In our Part III, we will have more to say about this.[165]

Thirdly, we stated that a kludge is not definable in terms of its neural implementation, even though changes in a specific set of behavioral and cognitive responses will primarily give reason to assume that the mechanism responsible for these responses has been modified such that a newly emerged kludge is part of it. However, given the prevalence of hierarchical and modular structure in explanatory mechanisms, we can expect that learning does involve a modification of precisely this structure. Indeed, the original neuroconstructivist account already contained a concrete conjecture regarding possible neural correlates of such a modification: "if the modularization thesis is correct, activation levels should initially be relatively distributed across the brain, and only with time (and this could be a short or relatively long time during infancy) would specific circuits always be activated in response to domain-specific inputs" (Karmiloff-Smith

---

[164] See note 160 above for the implicit learning theory, which in a similar way refers to representational redescription for explanations (Cleeremans 1997). Modelling implicit and explicit learning in order to account for empirical data on these learning processes, Sun and others are also interested in interaction between the two processes and their quite distinct task representations. Eventually, they argue for interaction between connectionist models for implicit processes and symbolic models for explicit processes, accounting for cases of inflexibility in implicit learning and of ineffectiveness of explicit re-learning of a given task (Sun, Slusarz et al. 2005).

[165] Indeed, research shows that differences in practice structure do not only lead to differences in skill learning, but also in the underlying neural processes. This can be explained in terms of the subjects' efforts to distill useful representations of the task at hand. Subjects learning a skill via variable practices appeared to engage higher level planning of the task with the involvement of prefrontal cortex, while subjects that practiced by mere repetition did recruit motor cortex only with less flexible control of the task (Kantak, Sullivan et al. 2010). Ethological observation in learning and imitating primates suggests that limited capabilities of representation at several hierarchical levels of a task and its components suffice to explain these processes and the shortcomings of primates in comparison to humans (Byrne and Russon 1998).

1992 5).[166] Even though this conjecture has received ample confirmation, it still leaves room for different specific neural implementations of this modularization process, as there can be several modifications of the responsible mechanism that can account for it.[167] In sum, notwithstanding the fact that modularization or increasingly focal neural activation patterns appear to be prime candidates for the neural implementation of kludge formation, it is not possible to identify kludge formation generally with a specific neural implementation.[168]

Is there indeed variation visible between individuals during the period leading to kludge formation regarding a particular cognitive or behavioral task, as our fourth characteristic implies? That is, should we expect identical changes in information processing or identical neural correlates to accompany the kludge formation process in different individuals, or is there room for variation even if the end result is comparable? The neuroconstructivist account suggests that although the final phase of mastering a particular task involves the individual's capability of handling multiple representations of the task simultaneously, it does allow individual differences in mode and timing of developing this capability (Karmiloff-Smith 1992). These differences are inevitable when one considers development and learning as the outcome of the dynamic interactions between systems, leaving intact the possibility that these interactions result in relatively stable cognitive and behavioral capabilities that are supported by developing modules (Elman, Bates et al. 1997). Indeed, gradual changes in a developing connectionist network can lead to abrupt and distinctive changes in cognition and behavior, in agreement with the modularization hypothesis of the neuroconstructivist approach.[169] In sum, we cannot but expect there to be relevant

---

[166] Below, in our discussion of the sixth kludge characteristic from the perspective of the modularization, we will emphasize how this perspective has difficulty with the idea of innate, neurally specified modules, as it is then hard to explain widespread consequences beyond a task domain of failures in the relevant modularization process (Karmiloff-Smith 1998).

[167] Meanwhile, brain imaging investigations in developing and learning subjects have been carried out that confirm the expectation, that: "developmental changes in patterns of brain activity appear to involve a shift from diffuse to more focal activation, likely representing a fine-tuning of relevant neural systems" (Durston and Casey 2006 2154). This confirmation does not only apply to infants but also to adults, as learning generally has been shown to be associated with an increasingly modular – and hierarchical – structure of the responsible neural mechanism (Bassett, Wymbs et al. 2011).

[168] It has been critically remarked that a wholesale denial of innately specified processes in the infant brain by some neuroconstructivists is at odds with some – although limited - neuroscientific evidence for such processes (Franck 2004). However, neuroconstructivist approaches do assign a role to some innate domain-specific predispositions – like attention biases - and even to some innate specifications which can be triggered by environmental stimuli, but argue that it is further interactions with the environment that subsequently determine the cognitive processes (Karmiloff-Smith 1992 ; Mareschal, Johnson et al. 2007)

[169] The observation of such abrupt changes has been described and underpinned with a tentative explanation in (Karmiloff-Smith 1992). More recently, a computational account that refers to bifurcation stages in developing networks has been added to such an explanation (Raijmakers 2007).

differences between individuals, at least during the intermediate stages of the formation of a kludge. These inter-individual differences can pertain to all aspects of kludge formation that were mentioned in the previous three characteristics – overlapping largely with the computational, algorithmic and neural implementation theories distinguished by Marr (Marr 1982).[170] This prevalence of inter-individual differences holds even if developmental processes generally yield increasingly hierarchical and modular structures.[171]

The fifth kludge characteristic refers to the meaning of its name, as a kludge is  not innate but derived from already present components – functions or processes, or a combination of both - with properties that are different from those of the eventual kludge. Here, the modularization account presents a nuanced answer that eventually supports the fact that re-use does not only occur with regard to neural areas but also to functional properties. Although modules allegedly are the result rather than the beginning of development and learning, these processes do not start from scratch in infants. Instead, a 'skeletal outline' of the brain that is present from the start, including several biases and predispositions, develops into an increasingly complex and modular structure, with changing connections to different brain processes (Karmiloff-Smith 1992 15, ff.). Nonetheless, an important difference exists between infant and adult brains, the latter containing more modularized structures than the former, corresponding with an increase in rather domain-specific cognitive processes (Karmiloff-Smith, Scerif et al. 2003). Associated with this is the fact that deficits existing at an early age, before extensive modularization has occurred, will have observable consequences in more than just a single cognitive domain because cognitive domains are only gradually isolated from each other: "the fact that domains are highly interrelated early in brain development (…) turns out to play a critical role in the formation of more general, albeit sometimes subtle, deficits in later development" (Karmiloff-Smith 2006a 47). In sum, modularization according to this account is indeed similar to kludge formation, modifying a mechanism that is responsible for a particular cognitive or behavioral task such that its weak modular structure with domain-general properties strengthens and its functional properties change, including its changing connectivity to mechanisms responsible for other functions.[172]

---

[170] Responding to the neuroconstructivist account presented in (Sirois, Spratling et al. 2008), Bateson adds to its interest in individual differences in development that it is not only the active role of the individual that matters but also its particular environment that influences development (Bateson 2008).
[171] Even though development generally builds on and further expands the hierarchical, modular structure of the brain, modifications of this structure depend so much upon interactions between different levels of the brain and its functions that a great deal of individual variability will be inevitable (Bassett and Gazzaniga 2011).

This latter aspect of changing connectivity brings us to the sixth characteristic: a kludge's involvement in further dynamic trajectories, its potentially being integrated as a component into a mechanism that underlies another function. For example, the observation that early visuospatial deficits in infants can have "cascading developmental effects over time on several *emerging* higher level linguistic and cognitive domains" (Karmiloff-Smith 2006a 47) confirms the occurrence of wide-spread neural 're-use' in the brain and the corresponding widespread consequences particularly of early functional deficits, as multiple functions will rely partly on shared neural components (Anderson 2010).[173] The question is whether such re-use also pertains to functions that have modularized due to development or learning, or whether modularization would preclude underlying mechanisms of such functions from being 're-used' again in other functions.[174]

Given Karmiloff-Smith's focus on developing brains and developmental disorders, this question is not a primary focus in her work, some of which concerns the wide-spread consequences of deficits in not-yet-modularized functions at later stages (Karmiloff-Smith 1998). And although the notion of modularization is attached to domain specificity, she does acknowledge that within any domain there are 'micro-domains' that function in turn as subcomponents – suggesting that micro-domain modules pertaining to gravity or pronoun acquisition can indeed be integrated in other, more comprehensive modules recruited for physics of language processing (Karmiloff-Smith 1992). Besides, although the prevalence of dissociations in adult brains suggests that modules are mostly domain-specific and segregated in adults, there are still domain-general disorders that rely on more focal deficits affecting a particular and modular function, as is the case with prefrontal deficits (Karmiloff-Smith 1992). Not just in the infant brain, therefore, but also in the adult brain we can observe that modularization can in fact facilitate the interaction of a modularized

---

[172] A more recent version of neuroconstructivism explicitly explains the development of complex representations on the basis of previously establish components, relying partly on modularization processes (Mareschal, Johnson et al. 2007)

[173] As noted in footnote 100 in part I, neural re-use is a prevalent phenomenon in the evolution and development of the brain. As a consequence, it was emphasized in that context that it is necessary to employ 'domain-neutral' terms to refer to component mechanisms that are being used in multiple cognitive and behavioral functions (Anderson 2010). Both issues concur with aspects of the presently discussed modularization account. Indeed, Anderson himself has acknowledged agreement  between his approach and the neuroconstructivist emphasis on modularization combined with interregional connectivity (Anderson 2008).

[174] It is worth noting that early connectionist models were 'highly task-specific and single-purpose', leaving unfulfilled Karmiloff-Smith's demand for developing models which go beyond modeling mastery of a specific task and that allow transfer of information across domains (Elman, Bates et al. 1997). It appears plausible that this limitation of these early models has contributed to rather limited interpretations of modularity, emphasizing informational encapsulation instead of allowing such information transfer, for example.

function with other functions, or its integration as a component in a more complex cognitive function.[175] Indeed, the twin processes of proceduralization and explicitation refer precisely to the increasing availability of a domain of expertise - which is being processed in a modularized way - to other processes. Thanks to the representational redescription which happens to such expertise during this process, its representational constraints render it available for both intra-domain and inter-domain relationships (Karmiloff-Smith 1992). Expertise, for example, can be held to rely on representational redescription yielding the availability of tacit or implicit knowledge to awareness (Feist 2013).[176] In Part III, we will discuss more explicitly how an agent with habituated intentional action patterns is better capable of flexibly responding to environmental information, as his action performance absorbs fewer resources than in the novice and leaves other resources available for processing this information and necessary action adjustments.[177] Several processes are involved in this development, of which the formation of relevant kludges is an important one, as we will see.

In this development of relevant kludges, the integration of environmental information plays an important role. The modularization account, in opposition to nativist accounts, not only emphasizes the importance of developmental and learning trajectories, but also the role of environmental information in these processes. It does acknowledge that an infant appears to be born with domain-specific predispositions

---

[175] Above we already noted that Fodor's notion of modularity should be supplemented with the notion of an assembly of modules, where component tasks rely on separate modules integrated in a complex assembly for the comprehensive task, as is defended in (Coltheart 1999).

[176] Expertise plays out differently in neural activations, as found in several fMRI experiments. Sometimes it appears that experts' brains show less activation in carrying out a task, sometimes experts recruit more neural areas. An explanation for this divergence in results could be that performance of a particular expert task must rely on the combination of a proceduralized and modularized component task with another cognitive task, while more focal activations suggest that this other task need only recruit a less complex mechanism. Indeed, not just a single form but various forms of the reconfiguration of – hierarchical and modular - neural networks are needed to account for both experimental and modeling results, depending among other things on task complexity (Bassett, Wymbs et al. 2011). Early imaging results suggested, for example, that an expert's skilled task performance allowed him to perform without much control and thus allowing him to again spend attention to information not directly relevant to the task, corresponding with increasingly widespread activations in comparison to novices (Raichle 1998). Imaging results for a motor sequence task in another experiment led to increasingly focal activation patterns for two distinct components of that task (synchronization and accuracy) and increasing connectivity between the neural mechanisms underlying these components (Steele and Penhune 2010). These examples show that it is highly task dependent how skill learning or expertise impacts on neural activation patterns and that such pattern changes can be multiple during the learning trajectory, but that in general such impact of learning is very common (Tracy, Flanders et al. 2003 ; van Mier, Tempel et al. 1998 ; Yin, Mulcare et al. 2009). Needless to say, the changing neural activation patterns eventually depend upon changes in synaptic activities at the single neuron level, which have indeed been found to respond rapidly to learning episodes (Xu, Yu et al. 2009).

[177] Cognitive flexibility as measured by drawing non-existent objects was shown to be greater in bilingual children, suggesting an interrepresentational flexibility that relies on representational redescription processes fostered by bilingualism (Adi-Japha, Berberich-Artzi et al. 2010).

that make it attentive and extra sensitive to specific environmental inputs, but it is the content of these inputs that partly determine further brain development and cognitive processes (Karmiloff-Smith 1992). Indeed, the importance of such interactions between environment and cognitive processes applies to most developing cognitive or behavioral functions, ranging from motor performance to drawing, reasoning about physics and symbol manipulation (Elman, Bates et al. 1997). Imprinting in chicks being considered a very simple example of such interaction (Elman, Bates et al. 1997), it is of limited relevance for our purposes compared to forms of learning that involve explicit instruction.[178]

The mastery of speech and symbol manipulation has an important impact on learning and associated processes. Research in bilingual children, for example, confirms that the mastery of more than just a single language provides these children with increased flexibility in cognition and behavior, probably depending on their being able to shift between two different representations of a single object or task (Adi-Japha, Berberich-Artzi et al. 2010 ; Bialystok, Shenfield et al. 2000). More than just generally facilitating learning and development via environmental inputs, these inputs provide subjects with extra resources to influence their other cognitive and behavioral processes.[179] For if development and learning rely in a crucial way on a series of Representational Redescriptions, then it should be possible to manipulate or reconfigure the representations involved such that they have specific influences on the learning processes and later outcomes.[180] Indeed, the acquired capability of explicitly manipulating and redescribing one's representations for a particular task yields many benefits for quick learning, correcting, adjusting, and cross-domain transfer of contents – and environmental information can play an important role in this capability (Clark and Karmiloff-Smith 1993).[181] Concurring with this, Hollis & Low conclude

---

[178] As noted earlier, imprinting in chicks has been interpreted as a demonstration that the strict distinction between innate and acquired constraints on cognitive processes is not useful. Though Wimsatt approached the matter differently from the modularization account, his account of the generative entrenchment of environmental information in a complex and modular system does correspond to a large extent with the former (Wimsatt 1986).

[179] However, such utility is only available at the developmental or learning stage where the relevant information has obtained the necessary representational status and not earlier (Karmiloff-Smith 1992).

[180] The authors of 'Neuroconstructivism' even describe a feed-back loop between a proactive child that affects its environment on the basis of its representations, eventually initiating further interactions that subsequently affect its own environmental inputs and thus its cognition and development (Sirois, Spratling et al. 2008).

[181] Language, symbols and narratives, for example, are being used for such manipulations and redescriptions. This will be discussed more specifically in chapters II.4 and III.4.

after their discussion of the effects of instruction and visual examples on children's drawings that: "[a] hierarchically organized exemplar training program may assist the progress of representational redescription by decomposing the task into an ordered series of subfeatures that partition the problem-solving space" (Hollis and Low 2005 641). Hence, external information not only speeds up learning but also influences the decomposition of the task's representation. As a result, the kludge formation that accounts for the behavioral mastery and flexibility pertaining to the task, involves the integration of environmental information. We will consider in Part III further examples of how such targeted analysis and manipulation of action representations by an agent can amount to his putting constraints on the space of actions available to him. Indeed, we will argue that expertise with particular intentional actions can imply kludge formation in underlying mechanisms, further sculpting the agent's space of actions – as when an opera singer's preferences in gender relations tacitly transpire in his performance of Don Giovanni.

## 2.3  Modularization considered as a process of kludge formation

After our comparison of the neuroconstructivist account of modularization with our seven characteristics of kludge formation, it seems not inappropriate to consider the modularization process as a form of kludge formation. Modularization is a process of increased encapsulation of a domain's – or subdomain's - information, being processed with an ever diminishing influence of other brain processes. The process is mainly observable via changes in cognitive and behavioral responses, with variability between stages and between subjects potentially occurring during this developmental process. This variability partly comes about through environmental information – like verbal instruction or examples – that plays a role during modularization and that can affect components of the task or the task as a whole. For such environmental information to be effective, a representational redescription of the relevant knowledge must already have obtained before in order to enable the subject to adjust its responses: the subject must already possess a degree of behavioral mastery before he can flexibly adjust to the new information. It is this complex interaction between the initial proceduralization of knowledge and its subsequent explicitation that is characteristic for this account and which offers some suggestions to our further development of the notion of a kludge.

First and most importantly, according to this account human subjects usually develop more than just a single representation of a particular domain of knowledge or behavior. These representations have different formats and accordingly also different properties. Some are more suitable for automatized motor performance while others, when redescription in a more abstract format has taken place, lend themselves better

for explicitation and flexible adaptation. The simultaneous availability of a plurality of representations pertaining to a particular task can explain surprising cognitive or behavioral results. This is the case when an agent shows temporary behavioral regress during a stage transition or when he can perform one and the same task according to different modes, each with different properties. In Part III, we will consider whether the same situation holds for the action determination processes on which we will be focusing our attention.

Second, given the presence of multiple representations after development and learning a task, it remains to be considered whether kludge formation can only obtain for the early stage of learning, or whether the more explicit representations that are developed at later stages of learning can also correspond with kludge formation. After reaching the stage where an agent can explicitly correct and adjust his task performance, may the adjusted representation in turn become proceduralized as well? Since neuroconstructivist research of modularization focuses mainly on children and development, the present account does not offer a clear position in this regard, although Karmiloff-Smith's account of herself learning to play the piano or solve Rubick's cube confirms that subsequent proceduralization can still occur in adults (Karmiloff-Smith 1992). This could theoretically lead to a situation where different implicit representations exist parallel to each other: one that emerged during the child's early learning and a second one that developed after explicitation, resulted again in an adjusted task representation.

With these two final remarks, we are already embarking on a discussion of so-called 'dual-process theories', which focus on the presence of two distinct types of processes underlying many cognitive and behavioral functions. Such processes are distinguishable in many respects, though they share largely overlapping domains of activation. Thus, as dual-process research demonstrates, it can occur that the two processes are provoked by the same environmental stimulus, thereby yielding two distinct behavioral responses to a single stimulus. This has led some authors to point out that humans have in fact 'two minds in one brain' (Evans 2003) which perform according to different types of processes, one of which is even labeled a 'cognitive monster' (Bargh 1999) as it performs its task more or less automatically. Particularly the latter phrase gives air to the negative assessment of this automatized type of processing, an assessment that is not common in the context of the proceduralization research, even though there are some similarities between the automatized and the proceduralized types of task performance. In the next chapters we will further explore the dual-process theories and consider whether they, too, in fact concern cases of kludge formation and thus bear witness to the capability of developing more than

one mode of performing a task once it is practiced regularly. If so, these theories may add some insight in the process of sculpting the space of actions as it happens with increasing expertise.

# 3 DUAL-PROCESS THEORIES AND A COMPETITION BETWEEN FORMS OF PROCESSING

When invited to sing a particular character on stage, it may be difficult for a lay and an expert singer alike to suppress the many performative and vocal characteristics that we automatically associate with a juvenile and virile Don Giovanni or with a senior and meditative Saint François. Indeed, it is highly plausible that without further thought or consideration, their performance of these characters will conform to certain stereotypes and biases that are commonly attached to a young womanizer or to a senior monk, as can be seen in their movements and gestures and heard in their vocal expressions. Directions aiming to portray the character differently and in such a way as to surprise the audience will at first require careful attention of the singers, in order to inhibit their usual performance and adjust it accordingly. If the singer has already mastered the relevant score, it will give him more freedom to enforce these adjustments, as singing the notes does no longer require as much attention. Only then may his Don Giovanni express some fear instead of bravura when inviting the Marble Guest to dinner, or his Saint François express less serenity and more vitality than is usually the case. After playing a role several times under different directors, an expert singer will in a sense have different schemas or Gestalts of that character available, lingering somewhere in his memory and awaiting complete or partial activation. Unfortunately, when participating in a particular production again after some time, the singer may notice that the reactivation is not without partially confusing the current directions with other characterizations of the role. Nonetheless, it will usually take less time to reactivate the desired performance than it took him to initially learn it.

The behavioral and cognitive stereotypes that are evoked in the previous section are not just suspected to flourish in everyday life, but have been demonstrated in various experimental situations as well. Typically, such experiments use specific words or images as primes, which are thought to strongly activate particular associations. These prominent associations are usually stereotypes and biases that subsequently modulate the cognitive and behavioral responses of subjects, even if they would explicitly reject such responses. For example, being primed with sentences or images referring to elderly persons, subjects tend to use longer reaction times to respond to tasks. When primes refer to politicians, subjects tend to be more verbose, and when referring to fast animals, reaction times were shorter.[182]

---

[182] See (Dijksterhuis and Bargh 2001) for a review of these and other similar experimental results.

What these experiments have shown, is that our cognitive and behavioral responses often demonstrate the influence of implicit cognitive processes without the involvement of conscious deliberation. This does not come as a surprise, given our discussion of the phenomenon of modularization in the previous chapter, where we argued that modularization entailed an example of kludge formation. Complex behavioral and cognitive functions like playing the piano or reasoning about physics, accordingly undergo a double featured process. For one thing, the progressive modularization that affects these functions makes them more encapsulated and thus less influentiable by other functions. Furthermore, the explicitation process associated with the redescription of the representations involved points in an opposite direction, potentially making relevant information available for exchange, correction, articulation or other interactions (Karmiloff-Smith 1992). Based upon this insight, the strict distinctions between declarative and procedural, or between conscious and unconscious, or between controlled and automatic forms of processing have been contested, since the developmental relation that connects them apparently does not lead to the simple substitution of one form of processing by another (Karmiloff-Smith 1992 26). This nuanced position is not shared by all researchers of human cognitive functions, as we will see in this chapter. This has to do with the recognition that our cognition and behavior are not always driven by a single process only.

Indeed, since the argument of this dissertation relies partly upon the recognition that in many cases there is more than just a single process available for performing functions - even performing a complex function like intentional action - and that this facilitates the fast and flexible performance of such functions, we will further explore the validity of this position. In chapter II.3 we will focus on the 'dual-process theories', which appear to be in opposition to the nuanced insight mentioned in the previous section.[183] As mentioned, these dual process theories generally claim that the human mind functions according to (at least) two distinct types of processes, sometimes leading to conflicting outcomes. One type is a 'cognitive monster' that can hardly be controlled and automatically determines cognitive and behavioral performances in a prejudiced and stereotypical way, in contrast to the consciously controlled type of processing that is normally assumed to determine agents' behavior (Bargh 1999).[184] This latter mode, however, has much less influence on human performance than is

---

[183] This may well be due to the neglect of developmental issues in research concerning dual process theories, which has been acknowledged as an 'unfortunate omission' in a recent review (Evans 2011).

[184] Notably, even Bargh later acknowledged that forms of self-regulation or control of automatic processes are possible, suggesting that the 'dichotomy' between the two needs reconsideration (Hassin, Bargh et al. 2009).

generally thought, as our stereotypical opera singer exemplified.

We should in this context not lose sight of the fact that we should be applying different levels of analysis, again, to what perhaps appears to be a relatively simple problem. In chapter I.3, we discussed the three levels of analysis or explanation that David Marr distinguished: the task or computational level, the algorithmic level which also includes the representations used, and the neural implementation level (Marr 1982). Without additional research, just looking at a cognitive or behavioral response alone will not inform us what information processing underlies this response, nor will we be able to plausibly argue in favor of a particular neural network that carries out the required processing. Accepting these distinguished levels of analysis, our subsequent discussion in chapter II.5 of mechanistic explanation further explored how we can investigate a cognitive or behavioral task by decomposing it in component tasks which are tentatively further decomposed and located in a complex and dynamic mechanism. The dynamics of an explanatory mechanism was argued to permit certain modifications to occur, usually corresponding with modified properties of the cognitive and behavioral performance for which the modifiable mechanism is responsible.

Reminding ourselves of this background is useful, as we are embarking on a discussion of a set of dual-process theories that claim that a seemingly identical task can be carried out via two – or more, in some models – different processes, with some theories also presenting hypotheses concerning neural systems that are allegedly responsible for these processes. The methodological considerations that were brought back to mind in the previous section intend to emphasize that there is no straightforward relation between tasks, processes and systems. This lack of strict correspondence goes in either direction, as was argued in Part I: neural systems are often re-used or re-cycled for more than just a single process and task, while a particular task can be executed with distinct processes, probably relying on correspondingly different systems. Since dual-process theories focus on the hinging role that processes play in connecting tasks to systems, a further remark seems in order.

Dual-process theories are concerned with the fact that certain forms of information processing are more complex than others and correspondingly rely on different neural systems. However, not all tasks can be performed by more than just a single type of processing, which makes such a task more likely to be constrained by a specific implementation of the appropriate processes. This seems to be the case for visual information processing, where binding different features of perceived objects together is limited by the constraints of the underlying complex system – memory being part of that system (Treisman 1998). Nonetheless, expertise plays a role in perceptual processes as well, research on chess masters having shown that they are capable of recognizing

complex board positions within seconds, relying on many stored positions in their memory (Gobet and Simon 1996). On top of expertise in the sense of sheer amount of experience, another form of expertise will be seen to play a role in determining the form of information processing that an individual employs in a certain task: his ability to engage one of several representational formats of the task at hand. Indeed, a characteristic of many tasks besides visual information processing is that they allow different representations and thus potentially also allow different neural systems to be involved when they are carried out (Halford, Wilson et al. 1998).[185] As was mentioned in the previous chapter on modularization, changing a representational format of a certain task can greatly reduce computational demands while enhancing the possibility for learning, correction and generalization (Clark and Karmiloff-Smith 1993).[186]

In this chapter on dual-process theories, we will also discuss whether it is possible that the performance of a particular task shifts from one form of processing to another form. Although we already know from the previous chapter on modularization that tasks as diverse as playing the piano and mathematical reasoning allow such a shift, peculiar to the present discussion is that in many occasions two distinct forms of processing appear to compete for determining the task outcome. Since particularly one of the two forms is considered to be seriously impeded by its large computational demands, this competition is often won by the 'cognitive monster' mentioned above, as this monster proceeds differently (Bargh 1999). Let us look more closely at the account given of the two forms of processing involved and subsequently investigate whether a shift of processing is possible, entailing another form of kludge formation.

## 3.1 Distinguishing between forms of processing, irrespective of tasks?

The dual-process theoretical assumption that a particular task can be carried out via two very different types of processing was based upon research like the experiment in which subjects were required to perform the so-called Wason-task, after which the researchers asked their subjects to "write down your reasons for choosing to examine or to ignore" a particular feature of that task (Wason and Evans 1975 142). The authors

---

[185] Halford et al. assert that visual information processing is highly modularized, making it very hard to 'reprogram' it in a strategic way as can be done with higher cognitive processes (Halford, Wilson et al. 1998).
[186] Cognitive complexity is proposed as reflecting "the ability to comprehend a cognitive domain with a variety of independent attributes for describing the objects in it" (Scott 1962 410). Cognitive flexibility is then defined as the ability to change the representations of the objects within the domain by focusing on different attributes, issuing in different decompositions of the domain. More specific is the notion of complexity that refers to entities and relations within a domain, where a domain usually allows multiple descriptions and descriptions at different levels of abstraction. For example, "Relations in a familiar domain can be more readily chunked, or higher order relations may be known that allow the structure to be represented hierarchically" (Halford, Wilson et al. 1998 811).

conclude from the incongruence between the observed results  and the introspective reports by the subjects that the performances and the introspective renderings of the task referred to different processes. Moreover, the biased reasoning that transpired in subjects' performance for the Wason-task appeared to be impenetrable or unavailable for their introspection. This rendered the introspective report delivered by the subjects post factum a mere construction or 'rationalization'. The researchers concluded from this observation that, contrary to common sense, at times a cognitive or behavioral task performance is not preceded by inferential reasoning, even though one would expect such reasoning to underlie the performance. On the contrary, such a performance is at times only followed by a rationalization which turns out to be erroneous and a mere construction, and should therefore not be taken to be reliable reports of subjects' actual cognitive processes  (Wason and Evans 1975).

Since those early investigations, dual-process theories have been proposed as explanations for the conflicting cognitive and behavioral responses demonstrated by subjects in the context of many different functions, ranging from social cognitive functions like attitudes, affect, self-regulation, social influences and blaming the victim (Chaiken and Trope 1999) to cognitive functions like reasoning and judgment (Evans 2008) and to forms of motor behavior (Hofmann, Friese et al. 2009). Indeed, even domain-unspecific functions like memory and learning are being approached from this dual-process theoretical perspective (Frankish and Evans 2009). Common to all such examples of dual-process theories is their emphasis upon those knowledge representations that are usually learnt implicitly and unconsciously and subsequently determine cognition and behavior in a similarly implicit and unconscious way, even though agents tend to think that their cognition and behavior is largely driven by explicit and conscious information processing.[187] Indeed, notwithstanding their differences, most dual-process theories share several attributes.

Although not all relevant authors agree that the two distinct types of processes are served by equally distinct – cognitive and neural - systems,[188] it has become common to refer to systems 1 and 2 respectively even when two types of processing are in fact

---

[187] Dual process theories are not unlike the reinterpretive work of the Masters of Suspicion – Marx, Nietzsche, Freud – who also did not accept the explicit and conscious self-accounts of fellow authors but instead argued that their mistaken or alienated self-construals in fact hid other factors determining human culture and thought (Ricoeur 1970). Not using the expression, they would probably agree to calling these other factors 'cognitive monsters'.

[188] Terminology among dual-process theorists is somewhat confusing and also liable to change. Two key authors – Evans and Stanovich - appear to agree in now favoring reference to two different 'types of processing' instead of two systems, with these processing types interacting when producing mental performances as well (Evans 2011). As this terminology concurs with our argument that a plurality of processes is available for many tasks for agents, who can learn to get some control of these, we will adopt this reference to 'types of processing'.

intended.[189] Similarly, most authors agree on the sets of attributes assigned to these two types of processing, notwithstanding some remaining differences between authors. In a recent historical and systematic review, the attributes of the two systems – or types of processing - are listed in Table 1, adapted from (Frankish and Evans 2009):

The previously mentioned distinction between task performance, form of algorithmic or information processing, and neural processes or systems is somewhat reflected in this table. The reference made to evolutionary age and distribution of the systems is particularly valid for their neural implementation, irrespective of the particular form of information processing carried out by those systems. As mentioned earlier, the fact that human and animal brains share many structures and properties does not withstand the fact that such structures are exapted for different forms of processing in humans, as well (Anderson 2010). Not surprisingly, the table especially pays attention to the differences in information processing between the two types, in

| System 1 | System 2 |
|---|---|
| Evolutionary old | Evolutionary recent |
| Unconscious, preconscious | Conscious |
| Shared with animals | Uniquely (distinctively) human |
| Implicit knowledge | Explicit knowledge |
| Automatic | Controlled |
| Fast | Slow |
| Parallel | Sequential |
| High capacity | Low capacity |
| Intuitive | Reflective |
| Contextualized | Abstract |
| Pragmatic | Logical |
| Associative | Rule-based |
| Independent of general intelligence | Linked to general intelligence |

**Table 1. Features attributed by various theorists to the two systems of cognition.**
Adapted from (Frankish and Evans 2009 15) with permission from the publisher.

---

[189] Once it is posited that two distinct neural systems underlie the two processes, a host of additional empirical hypotheses follow. For instance, dissociations between the two processes should be discernible in patients with lesions that affect one and not the other system. An early proposal for a two systems account assumed that the two processes recruit different memory systems, one being associative and the other rule-based (Smith and DeCoster 2000). More recently, neuroimaging results of experiments in which both processes are activated have led to the distinction between a reflexive, automatic system and a reflective, controlled system (Lieberman 2007).

other words to the 'algorithmic theory' in Marr's (Marr 1982) sense. The influence of rules or mere associations, the presence of abstract or contextualized information, the sequential or parallel nature of the informational process, and the difference in information load refer to these information processing differences. Associated with these processing differences are, finally, also observable differences in the task performance, such as its being conscious, its explicitability, its controllability and the influence of reflection on it.

Besides, and somewhat confusingly, Table 1 does not differentiate between attributes that refer to the learning process and others that refer to the activation of the previously learnt knowledge. This is most apparent with the pair 'fast – slow' which here refers not to the speed of learning but to the speed of activation, which happens fast for automatic and not for controlled processes. With regard to the speed of learning the pair would in fact have been the other way around, as conscious and rule-based learning can happen instantaneously, while unconscious and associative learning is dependent upon repeated exposure to the relevant information.[190] Finally, the table suggests that there are just two different systems or types of processes, while several authors argue that particularly system 1 or processing type 1 can be subdivided, with others arguing, conversely, that a single system underlies all different processing types depending upon the way it has been triggered by specific cues.[191] However, notwithstanding these differences there is agreement between most authors about two aspects, since: "[a]ll that really links dual process theories together is the nature of System 2 and the way in which implicit and automatic processes (of whatever kind) appear to compete with it for control of our behavior" (Evans 2006 205).

It is particularly the latter aspect that we will discuss in this chapter on dual process theories: the fight for control over cognition and behavior between the two

---

[190] This distinction between slow and fast learning has been aligned with distinctions between memory systems and types of content in the influential dual-process account presented in (Smith and DeCoster 2000). However, this content-based distinction between associative and rule-based learning has been challenged, as associations can also be taken to be a particular type of 'if-then' rules (Kruglanski and Orehek 2007). This has been defended even for the case of conditioning (Holyoak, Koh et al. 1989).

[191] Stanovich, for example, refers not to a single system 1 process but to 'TASS': The Autonomous Set of Systems. Common to these TASS is that they are fast, automatic and mandatory. However, he emphasizes that some TASS processes or particular goal states of these TASS processes may in some cases become automatic only after practice, which is not commonly attributed to System 1 processes (Stanovich 2005). Elsewhere, he distinguished the non-TASS processes in a 'reflective mind' and an 'algorithmic mind' the first signaling the need to employ non-TASS processes to a certain situation, the second then carrying out a reasoning task (Stanovich 2009). A different model distinguishes two automatic and two controlled processes. In this Quadruple model, the four processes may interact in a single task, depending on context, response tendencies, information availability and other task features. Consequently, the model allows the subject to engage in various forms of self-regulation and self-control (Sherman, Gawronski et al. 2008). In contrast, Kruglanski et al. propose a uni-model that responds differently to specific parameters of the task at hand (Kruglanski and Orehek 2007).

types of processing – which we will refer to from now on as automatic and controlled processing, respectively.[192] This fight for control is intimately linked to the distinctions in information processing together with the differences in the representations involved. One way of performing a task can involve a more comprehensive representation of the necessary information than another way of performing it. Associated with differences in representation format are, obviously, differences in information processing that rely on neural systems that can differ in kind and in number. The task of squaring the number 6, for example, is for many subjects a matter of activating their memorized table of 6, while others may have to add 6 sixes – which involves reliance on the capability to add and on working memory.

As we are more specifically interested in the process of kludge formation as it may contribute to a more stable 'sculpted space of actions', we will inquire to what extent is there a shift possible between the two: can a type 2 – controlled - process itself become automatic and thus gain more control over someone's cognition and behavior? Or is the line between the two types of processes strict and non-permeable, leaving automatic processes largely immune to the interference by a controlled process?

### 3.1.1 Considerations of the distinction between automatic and controlled processes

Before considering the possibility of automatization of type 2 processing, let us ward off the objection that type 2 processing can *per definitionem* not become automatized, as such a shift would render the conceptual distinction between automatic and controlled processes meaningless. There are at least three possible responses to such an objection. First, by making a gradual instead of a strict distinction between conscious and unconscious, or between implicit and explicit, or between automatic and controlled processing, we are better able to account for empirical and computational results. Notwithstanding the gradual nature of these distinctions, we can still recognize different phases with their own specific properties, for example with regard to the representations involved (Cleeremans and Jiménez 2002).[193] Second, as we noted in the previous chapter on modularization and the corresponding processes of proceduralization and explicitation, observable behavioral effects of

---

[192] Obviously, as is the case with a concept like 'implicit', the concept 'automatic' can be further decomposed in several features like unintentional, uncontrolled/uncontrollable, goal independent, autonomous, purely stimulus driven, unconscious, efficient, and fast. After analyzing these features, it is argued that the distinction with non-automaticity is gradual, rather than strict (Moors and De Houwer 2006), which concurs with our argument further below.

[193] Indeed, the previous section yielded the insight that learning – which was associated with a gradual yet multistage modularization process - can even result in: "the existence in the mind of multiple representations of similar knowledge at different levels of detail and explicitness" (Karmiloff-Smith 1992 22).

learning particularly concern the shift along precisely these gradients of a particular task (Karmiloff-Smith 1992). These effects, therefore, affirm that it makes sense to distinguish between these forms of processing, even though they are connected in a developmental or learning trajectory.[194] Third, specific to most dual-process theories is the assumption that a subject's actual cognition or behavior is the outcome of a competition between two different types of cognitive processing which differ particularly in the conditions of their activation.[195] Automaticity in this context does not so much refer to a strictly distinct type of processing but primarily to the activation of processes due to external triggers not necessarily selected by the subject. Controlled processing accordingly refers to the internal or intentional activation or – so we will argue – to the internal or intentional selection of those triggers that eventually activate cognitive processing.[196] Having considered these arguments against a strict conceptual separation of automatic from controlled processes, let us then proceed with the main question: what benefits should we expect to stem from a shift from controlled to automatic processing of a particular task? To answer this question, it is important to note the limitations that affect the causal or determinative power of controlled processes and to consider how these limitations are related to the kind of task-dependent information involved in these processes. Those limitations primarily concern the neural underpinnings of the processes that carry out the tasks and are therefore only in a derivative sense related to the information that is processed. As dual-process theories generally share the conviction that the process limitations

---

[194] In this context it may be noted that Aristotle's introduction of the dynamis-energeia gradient has provoked as much debate as it has helped thinking, particularly in the life sciences and medicine. The ancient resistance and the modern difficulty with similar notions may be related to a philosophical and – meanwhile – scientific preference for mathematics and physics with their more strict distinctions or fixed relations between entities, be they magnitudes or particles.

[195] The competition between processes can be configured differently. For example, two processes may simultaneously seek to control the outcome of processing, or one may seek to overcome or correct the preceding process, or a third process may select one of two processes to determine the outcome (Gilbert 1999). An account that has proven to be fruitful not only in explaining experimental but also in predicting results of training is a relatively simple model in which impulsive and reflective systems compete for the final determination of cognition or behavior (Hofmann, Friese et al. 2008 ; Strack and Deutsch 2004). It should be noted, however, that not all models assume such a competition and that other configurations are possible between the systems (Gilbert 1999). Interaction of systems, assumed to simultaneously contribute to a task performance, is assumed in accounts of (Cunningham, Zelazo et al. 2007 ; Smith and DeCoster 2000 ; Stenning and Lambalgen 2008 ; Sun, Slusarz et al. 2005)

[196] In contrast to the suggestion that controlled processing by definition cannot be automatized, one may thus doubt the value of a strict distinction between the two (Bargh and Ferguson 2000). Concurring with the latter is the observation that controlled, goal-directed attention allocation may influence subsequent automatic processing, which interplay of forms of attention: "may obfuscate the need for the distinction between automatic and controlled processing whatsoever" (Feldman Barrett, Tugade et al. 2004 567). Nonetheless, as is generally the case with terms that refer to functions or processes that are dynamically related, it may still be useful to distinguish them when there are empirically distinguishable properties involved.

matter most and that the limitations of controlled processing are responsible for the prominence of automatic processing, we will first approach that issue.

### 3.1.2 Processing limitations held responsible for the distinction between automatic and controlled processing

Note that the table discussed in the previous section lists as properties of controlled processes among others their being slow, their low capacity and their dependency upon consciousness (Frankish and Evans 2009). Indeed, it is such limitations in capacity of controlled processing, in view of the ongoing and multiple demands of a subject for immediate responses to his volatile environment, that would allegedly make some shift of the information processing load to automatic processing desirable. As automatic processing is fast, has greater capacity, can occur in parallel and does not need to involve consciousness, it is considered to be the default type of processing, leaving to controlled processing a limited role which can only be deployed sparingly.[197] Obviously, once controlled processing of a certain task could also become automatized, task performance would perhaps no longer be subject to the limitations that hold for controlled processes. In our discussion of the dual-process theories we will argue – particularly in section II.3.1.5 - that the complexity of a particular task should not be considered as a static fact. Instead, there are strategies available for reducing the complexity of the information that requires processing for a particular task, comparable to the process of Representational Redescription discussed in section II.2.1, that is involved in children's mastery of certain skills. Consequently, not only is complexity to some extent adaptable, it is also problematic to determine capacity limitations in terms of information processing limitations. Nonetheless, we will consider some dual-process accounts that focus on a particular processing factor or neural component as being responsible for such limitations.

Authors do not completely agree in their diagnosis of the most important bottleneck that yields this limited capacity of controlled processing. This disagreement may be partly due to differences in focus on one or another component of a complex and dynamic mechanism for information processing and decision making, where this mechanism may not at all times engage all of its components with equal strength. However, as both processing types are involved in information processing, the capacity limitation at stake makes itself felt particularly with regard to the *quantity*

---

[197] Considering different prominent obstacles for a dominance of controlled processing – Reflective system processing, in their terms – like time pressure, cognitive overload and alcohol intoxication, Hofmann et al. conclude that all of these obstacles are related to impediments of working memory (Hofmann, Friese et al. 2009).

of information that is being processed. Where is this informational bottleneck to be located in the mechanism – and is there a single bottleneck, or are there potentially several components involved?

Some authors identify consciousness as the bottleneck. As was visible in Table 1 above, most dual process theories consider the type 2 processes to be conscious. The 'unconscious-thought theory' appears to locate the shortcomings of conscious thought in dealing with complex problems in its limited capacity for dealing with large amounts of information (Dijksterhuis and Nordgren 2006).[198] Nonetheless, consciousness is here not defined in such a way that it clarifies why consciousness should present the bottleneck nor how we should determine its capacity.[199] Instead, the unconscious-thought theory refers to Miller's account of the informational limit that has been so influential since his article on "The magical number seven, plus or minus two" (Miller 1956). Interestingly, that account focused on the capacity of memory, and not on the capacity of consciousness.[200] Indeed, it is questionable whether consciousness can be held responsible for the capacity limitations.[201] Perhaps, therefore, this alleged limited capacity of consciousness could be a downstream effect of memory – both being part of a more comprehensive information processing mechanism. Let us first consider whether we can explain differences in cognitive and behavioral responses, following the suggestion that different systems are involved in types of information processing, like different memory systems.

---

[198] Unconscious-thought theory is usually distinguished from dual-process theories in that it does not assume the presence of two different systems – though this is not a strict prerequisite for dual-process theories, as we observed above. Furthermore, the theory differs from common dual-process accounts in that it does not criticize but highlights the value and optimal outcomes of certain effortless, unconscious thought processes. Crucial, however, is its denial of a strict distinction between conscious and unconscious thought with respect to the kind of input – associative or rule-based – that is used (Dijksterhuis and Nordgren 2006). Our account of a shift from controlled to automatic processing concurs with that denial.

[199] In another article, conscious thought is defined as 'deliberation-with-attention', referring now to attention as a limiting component (Dijksterhuis, Bos et al. 2006). The authors there suggest that doing arithmetic requires conscious attention, failing to acknowledge that doing arithmetic consciously and attentively yields extremely divergent results in novices and masters, even when applying identical arithmetic rules.

[200] Miller underscores the importance of recoding – particularly linguistic recoding – of information that helps humans to counter to some extent the limitations of memory (Miller 1956). This is not denied in the unconscious-thought theory, even though consciousness is being denied a functional role with regard to determining the productive information processes via encoding and recoding of information (Dijksterhuis 2007).

[201] Such an association of consciousness with computational capacity limitations is explicitly denied in a version of the 'workspace' theory of consciousness, which equally rejects the notion that consciousness corresponds with particular neural systems. Instead, the authors argue that it is the specific type of neural activation that is responsible for making a particular content consciously available (Dehaene and Naccache 2001). The workspace theory of consciousness even allows specialized computational processes to run unconsciously in parallel, each being responsible for making only relevant parts of the overall task consciously available, while conscious processing occurs serially at the same time (Shanahan and Baars 2005).

### 3.1.3 Memory systems invoked for the explanation of the distinction between automatic and controlled processing

Indeed, a distinction can be made between two different memory systems, which are also involved in different forms of learning: "a memory system that supports gradual or incremental learning and is involved in the acquisition of habits and skills and a system that supports rapid one-trial learning and is necessary for forming memories that represent specific situations and episodes" (Sherry and Schacter 1987 446). It is the latter memory system, playing a role in many cognitive functions as working memory - its central executive component[202] in particular – and its limitations that essentially impedes controlled processing. Many of the differences that have been found in individuals with regard to the interaction between the two types of processing may be attributed to differences regarding their working memory.[203] As a result, an individual can be called either a 'motivated tactician' with high working memory and consequently large controlled processing capacity, or a 'cognitive miser' with low capacity and therefore with a relatively larger role for automatic processing (Feldman Barrett, Tugade et al. 2004). The question then presents itself whether subjects can only demonstrate cognitive and behavioral responses according to either one of these types, or whether they can to some extent shift between these two types.

Several authors focus on the role of memory for answering this question on shifting between processing types as a way to modify responses. The assumption is that subjects need to sculpt their space of actions by bringing about such a shift in processing. According to the so-called Reflective-Impulsive model, Type 1 or automatic processing is carried out by the impulsive system which 'slowly and gradually' yields a network of associative connections related to behavioral schemas. The result can be considered to be a 'conceptual and procedural long-term memory' that can process large amounts of information in parallel, with activation of multiple response options as a result. In contrast, Type 2 or controlled processing activates behavioral schemas by way of the reflective system. Controlled processing relies on executive control and working memory and is accordingly impeded by severe capacity limitations (Deutsch and Strack

---

[202] In the words of a pioneer of working memory research, the central executive is still: "the most important but least understood component of working memory" (Baddeley 2003 835). Nonetheless, in his review Baddeley associates working memory with forms of self-control and regulation and contrasts this with implicit forms of control which do not rely on working memory – concurring with the dual process approach discussed here.

[203] However, it would be mistaken to assume that automatic processing which issues in a response does not engage working memory, just like it is a mistake to assume that controlled processing in no way involves automatic determination of components of the response – like in implementation intentions (see more on the latter below) (Bargh and Ferguson 2000). So even though working memory limitations may make themselves more notably felt in controlled processing, working memory is involved in automatic processing as well.

2006). According to the model, both types of processing activate in their separate ways behavioral schemas, and a competition between these schemas eventually leads to the execution of a particular behavioral schema.[204]

However, even though the distinction between controlled and automatic processing is perhaps associated with differences in the recruited memory systems, it cannot be associated with a strict distinction between tasks. That is to say, one and the same task can be performed according to different strategies, relying on different task representations and as a result, can also be associated with different – neural – systems. For example, applying the Reflective-Impulsive model, research has demonstrated that with repeated conscious and intentional exercise of a particular behavioral schema, this schema will over time be processed with the impulsive instead of the reflective system. As a result, preferred and self-regulated behavior can become part of automatic processing, thus avoiding the limitations of controlled processing (Hofmann, Friese et al. 2008). Indeed, in contrast to the sometimes worrisome reflections on the prevalence of biased and stereotypical responses that are raised in the context of dual-process theories, room is left open for a process of 'sculpting the space of actions' even in this theoretical context.

### 3.1.4 Some strategies that allow a shift between automatic and controlled processing

The repeated practical exercise of self-regulated behavior as a way of modifying the mechanism responsible for our behavior or cognition generally implies a specific role for controlled processing. However, this is not the only strategy available to agents when they seek to change their behavioral or cognitive response patterns, upon the recognition of their limitations of controlled processing. Other strategies are available, differing in nature and in the focus on the action component needing adjustment. This section will mention a couple of such strategies, merely to show the variety and to demonstrate that kludge formation can affect responsible mechanisms.

To begin with, automatized self-regulation need not just pertain to behavioral properties, like the goal of an action. It can also be directed at emotions associated with actions. Several strategies have been shown to be effective in such automatic self-regulation. For example, an agent can prepare himself by formulating so-called implementation intentions that concern a particular future action in a specific

---

[204] A different approach to the distinction between subjects' accustomed behavior and their goal-directed actions uses the construct of habit as environmentally cued behavior. This construct, too, recognizes the interaction between the two processes which enables agents to try to mitigate undesired habits (Wood and Neal 2007).

situation, which are shown to influence his future actions even if these are done automatically (Gollwitzer and Brandstatter 1997). In a similar vein can an agent prepare himself by articulating emotions that he considers to be appropriate or he can practice to automatically withdraw his attention from negative stimuli or to avoid negative interpretation biases (Gyurak, Gross et al. 2011).

More common are strategies that are directed at the outcome or goal of an action. Apart from practice-based modification, an agent can also activate consciously and intentionally a preferred behavioral schema with controlled processing and thus prepare for future situations by alleviating the taxation of controlled processing during those situations. Such a strategy makes use of agents' capability of automatic goal-pursuit, relying on the memorized representation of a habituated goal-directed action, which can be activated not only consciously but also without awareness (Bargh, Gollwitzer et al. 2001).[205]

For such unconscious goal-directed responses to occur, an agent can engage in the explicit formulation of intentions for specific goal-directed behavior under particular conditions. With such implementation intentions, an agent can obtain results that are comparable to habituated responses, even though the two strategies differ (Aarts and Dijksterhuis 2000). With the articulation of explicit 'if-then' rules for action, agents can anticipate future situations such that they act more reliably in such situations according to their previously formulated intentions. Yet they respond to a present situation with the intended behavioral response in an immediate and efficient way, without the involvement of conscious intent (Gollwitzer and Sheeran 2006).[206] Indeed, what has been formulated via a controlled process has become automatized such that the intended behavioral response will be automatically activated by the anticipated situational cues (Webb and Sheeran 2007).[207]

While the implementation intention strategy relies on the preliminary activation of a behavioral schema, thus facilitating its execution at a later stage, an alternative for agents is to engage in counterfactual thought as a self-regulation strategy in order to

---

[205] In spite of earlier convictions that automatic goal pursuit is by nature inflexible and irresponsive to altered environmental or internal states, more recent research has demonstrated that automatic goal pursuit can under circumstances in fact be flexible and responsive (Hassin, Bargh et al. 2009). Agent's preparation can play a role in this, as can the representation of the task, as we will argue below.

[206] These three features, immediacy, efficiency and lack of conscious intent, are considered prime attributes of automatic processes, as the authors note correctly (Gollwitzer and Sheeran 2006)

[207] Indeed, one of Kruglanski's reasons for developing a uni-model in which the dual processes are unified is that the associations that determine automatic processing can also be considered as if-then rules – even though they are not explicitly and consciously formulated by an agent during his response (Kruglanski and Orehek 2007). Lieberman makes a similar admission and has an equal dislike of the strict separation of automatic from controlled processing, inspiring him to the development of yet another model in which interactions between these processes are prevalent (Lieberman, Gaunt et al. 2002).

diminish the chance of executing an undesired action. Such simulation or imagination of future, hypothetical response options is typically associated with controlled processing (Evans 2008). In this case, remarkably, activation of a response option through controlled processing influences subsequent automatic processing under specific conditions. For example, counterfactual thought, devoted to an alternative course of action or judgment while considering a past event, can be used to prime a future cognitive or behavioral response and leads to less biased responses (Galinsky and Moskowitz 2000). Consisting of the activation of memorized representations of previous events and then on the simulation of alternatives – or 'variations on a theme' -, counterfactual thought depends on cognitive and neural processes that are similar to those involved in normal action planning and coordination. (Narayanan 2009).[208]

Now that we have considered some different strategies for automatic self-regulation, a few remarks should be made. First, these strategies and the shift from controlled to automatic processing that they bring about, once again emphasize that it is implausible to strictly separate the two types of processing. Indeed, one of the reason for proposing the Quad-model as an alternative, is to account for the various forms and focuses of self-regulatory strategies that agents have at their disposal (Sherman, Gawronski et al. 2008). Second, many strategies in some way involve a representation of a relevant action or action situation.[209] Such a representation can refer to an action at different levels of grain or abstraction, with different regulatory outcomes. Given the relevance of this aspect of representation, which will be more at the focus of Part III, let us pause for a moment to consider such task representations, or the 'algorithmic' theories (Marr 1982) involved.

### 3.1.5 Representational differences and the shift between automatic and controlled processing

To summarize the previous section, both the intentional adaptation of a practice-based

---

[208] Counterfactual thought allegedly exploits the representation of 'Structured Event Complexes', which generally have a hierarchical structure and a temporal sequence. Based upon the analysis of these SEC's, counterfactual thought is assumed to be most effective when focusing on the central dimensions of action-inaction, self-other, and event outcomes (Barbey, Krueger et al. 2009).

[209] There is a heated debate about the notion of representation in cognitive neuroscience. The debate concerns questions such as whether we need to involve representations at all, or should instead employ notions from dynamical systems theory, for example (Keijzer 2002). A related topic is whether the notion of representation adequately captures the properties of distributed information in neural networks, influencing network activations. (Bechtel 1998) defends this position against among others (van Gelder 1998). Given the fact that self-regulation can be successful with the preliminary invocation of an action representation, we consider the notion both plausible and useful here. This concurs with the observation that representations are involved in forms of learning, transfer and correction of cognition and behavior (Clark and Karmiloff-Smith 1993), as is also the case in so-called 'representation-hungry' tasks involving distal, non-existent or highly abstract action properties (Clark and Toribio 1994).

strategy and the strategies relying on the explicit formulation of specified intentions or of counterfactual intentions, therefore, are capable of modifying the agent's mechanism that is responsible for his behavioral response in particular situations. Obviously, the modifications obtained with these strategies are not equally structural or long-lasting. Indeed, further down – in chapter II.4 - we will discuss further strategies or tools available for such modifications that involve instruments or representations, linguistic and otherwise, that are even external to the skull in which our information processing takes place. Nonetheless, such modifications of 'internally' represented information do result in some change of an agent's sculpted space of actions, corresponding with some shift from controlled to automatic processing. This shift can occur for actions that conform to a more or less specified representation of a cognitive or behavioral response. As a result, the limitations in capacity of automatic processing are circumvented, notably the limitation of working memory that is crucial for controlled processing. Indeed, having presented his influential account of the limitations of memory, Miller emphasized that a recoding of information can help to mitigate these limitations (Miller 1956). Fortunately, therefore, the distinction between the two processing types is neither strict, nor static, allowing a specific controlled process to shift to automatic processing, thus alleviating the burden on working memory.

In the next section we will close our discussion of dual-process theories with a consideration of the seven kludge characteristics, which we use as a means to estimate the agreement of these theories with the observation that modification of an explanatory mechanism responsible for an agent's action can occur. Before engaging with this evaluative task, we'll devote this section to a discussion of the task or information processing involved in both automatic and controlled cognitive processes. We already know from our earlier discussion of the modularization process in development and learning that the representation involved in certain cognitive or behavioral tasks does matter in an important way. Both the proceduralization and the explicitation that were found to be associated with the phase-wise mastery of a cognitive arithmetic task or with a behavioral task like playing the piano were argued to rely on a process of 'Representational Redescription' (Karmiloff-Smith 1992). More generally, the human capability of employing different formats of representation for the performance of a certain task enables humans to engage in corrections, generalizations, and transfer more than animals do (Clark and Karmiloff-Smith 1993).

How can we extend this insight to the present issue of two allegedly distinct, yet competing processes that are potentially involved in performing the same task? As we noted regarding Table 1 on the two cognitive processes in section II.3.1, some of the attributes included there refer to the differences in the representations involved in

those processes. However, these attributes should not seduce us into thinking that a strict distinction is at stake, although the table associates the sequential processing of rules and abstract information with controlled processing while leaving for automatic processing the parallel processing of associations of contextualized information (Frankish and Evans 2009). Instead, we will present a couple of arguments why such a strict distinction between the tasks or representations involved must be rejected.

To begin with, in contrast to the distinction between associations and rules mentioned in the table, associations can be considered to be a particular type of 'if-then' rule (Kruglanski and Orehek 2007). Indeed, even classical conditioning of behavior, typically depending on association learning, has been defended as being a matter of learning to follow such a rule (Holyoak, Koh et al. 1989).[210] Comparable with this identification of association learning with a kind of rule-learning is also the 'Representational Redescription' process referred to above, when more abstract re-representations can arise spontaneously after repeatedly performing a task.[211] Such a process can occur as a consequence of performing embodied actions, resulting in rule-learning in children and yielding observable results without involvement of verbal, conscious and controlled processing (Boncoddo, Dixon et al. 2010). Again, as argued earlier, controlled and automatic processing should not be placed strictly apart, nor should we assume the information involved in these types of processing to be strictly different. Let us pursue this last point somewhat further.

Defining the complexity or computational load involved in information processing has turned out to be an intricate problem, which we will not try to solve here.[212] Our more modest aim here is to argue that – irrespective of its specific definition - the complexity of a task is not a static given but allows some adjustment. For example, a particular approach to this issue starts from the analysis of complexity as *relational complexity*: not just the number of items implied in a certain task matters, but more specifically the number of arguments or relations between them (Halford, Wilson et al. 1998).[213] Complexity then increases with the number of interacting variables, but

---

[210] It is usually held that relatively lower level cognitive processes, like visual perception, are modifiable merely through learning simple association rules. In contrast to this, research demonstrates that expertise with rather abstract mathematical reasoning does also influence early perceptual processing, as experts visually group notational elements in mathematical exercises different from novices (Goldstone, Landy et al. 2010). We will return to this in section II.4.2.1 below.

[211] Such learning could be described as self-organization of the relevant neural networks, enabling the emergence of new properties – in this case, new representations (Stephen, Dixon et al. 2009).

[212] There are several different definitions of complexity available, for example with regard to the question whether or not evolution leads to increasing complexity (Chu 2008). Preliminarily, however, one would need to decide what the valid comparisons are in this context, what time scales are to be applied, and how the complexity is operationalized and measured – which is difficult without a definition (McShea 1991).

[213] A different yet comparable account focuses on the number and the complexity of the relations that are used in task performances and is equally interested in subjects' flexibility in using a different set of relations to reduce a task's complexity (Zelazo and Frye 1998)

especially when the number of different relations between these variables increases. When a series of even numbers needs to be added together, the complexity is less than when addition and multiplication must be alternated, for example.

However, such relational complexity can be transformed. Consequently, however it is quantified, complexity is not static and subjects can employ different strategies to reduce the complexity of information involved in a particular task. Such strategy change generally modifies a subject's performance and can yield different benefits, as each strategy activates different representations.[214]

One strategy is to segment a complex task in smaller component tasks, which can then be processed serially, another is to reduce complexity by chunking or collapsing the information into other relations. In that case, a hierarchical structure of relations is developed. The latter strategy is dependent, obviously, on two types of familiarity: "Relations in a familiar domain can be more readily chunked, or higher order relations may be known that allow the structure to be represented hierarchically" (Halford, Wilson et al. 1998 811). Our argument is not that with such strategies all processing limitations can be overcome, but merely that the task complexity can be reduced and consequently processing can be facilitated. Besides, we should remind ourselves what we learned above, that a redescription of representation can lead to loss of some information, as it often involves an abstraction of information. Or it might involve an exchange of information: for example discarding some information that is apparently irrelevant for the task at hand in favor of a different yet simpler representation.[215]

Indeed, other strands of research have confirmed that cognitive processes generally represent information at different levels of abstraction. In such cases, chunking may be a useful strategy. For example, the perception of unfamiliar actions by children already involves such hierarchical encoding, as they encode a complex action at several levels of abstraction simultaneously – encoding both a complex action and the component actions which make it up (Baldwin, Andersson et al. 2008). Indeed, subjects automatically segment perceived environments and actions according to a hierarchical

---

[214] According to this approach: "working memory is the workspace where relational representations are constructed and it is influence by knowledge stored in semantic memory" (Halford, Wilson et al. 2010 499).
[215] One critique of Halford's et al. target article argues that 'skill theory' provides a different account of learning to process higher relational complexity by: "(a) learning two simpler skills or networks, (b) mastery to allow parallel sustaining of the two, and (c) coordinating them in a new relation through multiple steps specified by skill/network combination rules" (Coch and Fischer 1998 835). The nature of the new relation requires some scrutiny. Not dissimilar is the critique from Cognitive Complexity and Control theory, which also emphasizes that some relation types may be more difficult to master than others, predicting differences in developmental trajectories dependent upon the relation types involved (Frye and Zelazo 1998). Notwithstanding their criticism, these authors still concur with the importance and prevalence of different complexity reduction strategies.

structure. In such cases, prior conceptualization of an action to be observed does not always determine the subsequent encoding of the observed action (Zacks, Kumar et al. 2009). Nonetheless, dependent upon a task to be performed, for example imitation, subjects are capable of adjusting the level of encoding of these perceptions to the task at hand. Verbal description can facilitate such encoding at several levels of hierarchy, but such differentiated encoding occurs as well after repeated observation (Hard, Lozano et al. 2006). The fact that such hierarchical encoding is not dependent upon conscious and language-dependent processing is supported by the fact that with their imitation learning capabilities, primates can be seen to engage in it as well (Byrne and Russon 1998). Here again, familiarity with a domain and with pertinent rules or hierarchical representations facilitate cognitive processing.

Before summarizing the present discussion, we need to engage with a related issue which can be illustrated with our example of expert singing. We emphasized that an expert singer is capable of modifying his performance in accordance with specific directions or intentions. Apparently, so we argued, an expert not only has flexible control over his performance, he can also access his performance via verbal or other cognitive strategies. This seems to contradict an alternative account of expertise (Dreyfus and Dreyfus 1986), which has been developed further since. According to this account, skillful performance does not rely on representations at all, implying that a skilled performer responds to environmental information without the invocation of explicit representations: "Unlike deliberate action, skillful coping turns out to have a world-to-mind direction of causation" (Dreyfus 2002 380) which allegedly leaves no room for representations of the world that are stored somehow in the mind.[216] Indeed, where a person engaged in improving his performance may rely on analytic reasoning, once skilled, this person is held to rely on intuitive decision making, which apparently rules out the contribution of representations (Dreyfus 2004). The implicit and automatic way in which an expert responds to his environment while skillfully coping with it, is taken to imply that representations no longer play a role in his cognition or behavior.[217] The shift from explicit and controlled performance of a task

---

[216] Dreyfus' phenomenological critique of cognitive science is connected to his phenomenologically inspired critique of the important role it assigns to representations. In that, however, he differs from Wheeler, who equally aims to develop a phenomenological account of cognition and behavior that sits well with cognitive science, yet leaves a role for representations. Importantly, these representations need to bear relevance for the agent, according to his account (Wheeler 2010)

[217] An ethical implication of this account of skill acquisition developed by Hubert and Stuart Dreyfus is that experts cannot be required to articulate and explicate their intuitive decisions, as they don't make these along the analytical, reasoning processes as novices do (Selinger and Crease 2002). This alone seems a very problematic aspect of the account, even though it may not be an inevitable implication. Holding experts morally responsible for their actions requires them to be capable of offering representations of these. Given our inclination to demand this of experts, it seems that our moral intuition does not concur with the account of intuition that the Dreyfuses provide.

to an implicit and automatic – in sum: intuitive – performance would correspond with a more holistic and non-analytic approach to a situation which must be explained in terms of neural dynamical systems and not in terms of representations, this account holds (Dreyfus 2009).

Although we risk reducing this discussion only to the potential role of representations, there are a few aspects worth mentioning in response to this account. First, in a critical response to this intuition-based account of expert skill, it has been pointed out that there is a lack of empirical evidence for the five distinguished stages and for the absence of analytic reasoning in expert performance (Gobet and Chassy 2009). Second, once we give an account of expert skill that involves an important role for the chunking of information, it implies that an expert is capable of capturing and processing strongly associated pieces of information at once. This is different from simply listing details as novices may need to do, simultaneously offering a way to account for an expert's difficulty of articulating his expertise (Gobet, Lane et al. 2001). This would require the articulation of chunked information that is still a representation, though it has meanwhile become a redescribed or re-represented one. Both neuroscientific and simulation studies suggest that this chunking process is even more complex than mentioned above. Indeed, chunks appear to have some hierarchical structure, as they function as templates with slots being left open for variable environmental information. As such, it can explain not only that experts are very quick in responding to complex environmental information – like in playing chess or in nursing – but that they do so while simultaneously responding flexibly to changes in detail of the environment (Gobet and Chassy 2009). Third, such chunks can become associated with emotional features, explaining why experts often respond with strong affective tendencies to a certain situation (Chassy and Gobet 2011).[218] Finally, without an account of expertise that involves the representation of information, it will be problematic to account for experts' capability of quickly learning, correcting, modifying, transferring and generalizing specific performances of their skill (Clark and Karmiloff-Smith 1993). Indeed, in our example of expert singers who can modify their performance of a Don Giovanni or other character when receiving explicit instructions, we can witness how representations play a crucial and effective role even when an expert usually relies on implicit and automatic processes for his performance.

After this short excursion, let us summarize the previous arguments. We have argued that, not just on the basis of considerations regarding automatic and controlled

---

[218] Building upon Dreyfus' account of expertise and adding to it the notion of 'action readiness' that stems from Frijda's theory of emotions (Frijda 1986), Rietveld has argued that an agent's concern-full, unreflective action is partly determined by the affective response that environmental affordances evoke (Rietveld 2008).

processes or their underlying systems, but also based upon the consideration of the information or representations involved, a strict distinction between controlled and automatic processes is not warranted. Instead, interactions between the processes should be expected. Indeed, a neural network model containing different levels of both automatic and controlled processing and assuming an iterative interaction between the two types of processing, turns out to be plausible both regarding system requirements and in predicting behavioral results (Cunningham, Zelazo et al. 2007).[219] Therefore, apart from the fact that shifts of processing do occur in various ways, there are also other reasons for not keeping the two types of processing strictly apart.

It is now time to wrap up the observations regarding the two types of processing by considering whether there is kludge formation at stake in this context, too. For we are interested in this part in modifications of an agent's mechanism such that we can observe some changes associated with his performance of a task, while other aspects remain the same.

## 3.2 Automatization of controlled self-regulation and the seven kludge characteristics

A kludge is assumed to be a part of an explanatory mechanism, being involved in the modification of such a mechanism. Indeed, that modification is to a large extent due to the formation of a new kludge, which obviously has some consequences for the mechanism, its properties and its observable behavior. In this section, we consider whether we can apply the notion of a kludge to cases in which a shift occurs from controlled to automatic processing. When development or learning has led to the establishment of a kludge, its first characteristic – presented in section II.1.1 above - reads that it should be recognizable on the basis of functional rather than other properties. If we refer once more to Table 1 included in section II.3.1, listing attributes pertaining to controlled and automatic processing, it clearly present mainly functional differences between the two, like speed, capacity, efficiency, explicitness, involvement of consciousness, contents, and so on (Frankish and Evans 2009). Indeed, comparable models, distinguishing between reflective and impulsive (Strack and Deutsch 2004) or between reflective and reflexive systems (Lieberman 2007) agree in that they distinguish between types of processing and make similar functional distinctions between the two. In the case of a shift between types of processing as a result of automatized self-regulation, for example, we can observe this primarily on the basis of the response

---

[219] A similar – bidirectional - interaction between the two types of processing can also account for social responses, depending upon the excitation or inhibition of specific representations (Adolphs 2009). See note 195 above for additional comments on different configurations between the types of processing or systems.

properties. Similarly, an incomplete shift or the involvement of both processing types is determined equally on the basis of observable cognitive and behavioral responses (Rydell, McConnell et al. 2006).

Second, recognizing a kludge on the basis of properties of cognitive or behavioral responses in which it is involved – as part of the responsible explanatory mechanism – still does not enable us to derive from these properties a specific algorithmic theory that can account for its formation. In section II.3.1.4 we discussed several strategies that could be employed for self-regulation, all involving a shift from controlled to automatic processing. This is clearly visible when comparing implementation intentions (Gollwitzer and Brandstatter 1997) with counterfactual thought (Galinsky and Moskowitz 2000): the first strategy primes or activates a particular response or components of it, whilst the second strategy does partly work through de-biasing or de-activating a particular representation instead. Still other strategies are possible as well, that do not so much focus on explicit and language-based representations, as aim to manipulate an agent's attention bias or behavioral approach tendency instead (Stacy and Wiers 2010). Moreover, in the previous section we noted that some reduction of the complexity of representations that are being processed can be obtained through their segmentation or chunking (Halford, Wilson et al. 1998). Consequently, a shift in processing could then occur. Common to all such strategies, however, is that they involve a form of self-regulation without constant conscious control. We are suggesting to consider this relatively stable adjustment of automatic processing as a result of kludge formation, such that the responsible mechanism yields more appropriate and self-regulated responses than before. It has to be acknowledged that the efficacy and durability of the established kludge can differ, since the mechanism yielding an agent's response can still operate with or without involvement of this kludge when activated in a particular situation.

Irrespective of the loose – not strict – independence of the neural implementation theory from the accounts of both the task at hand as well as its strategy that was defended in Part I, it is still relevant to consider the neural implementation of the kludge. However, given the multiple reasons given in section II.3.1.1 against the plausibility of a strict distinction between automatic and controlled processing, identifying neural correlates for the process of kludge formation may be complicated, too. Moreover, although there may be some agreement between authors about the nature and neural implementation of controlled processing – involving several PFC areas[220]-

---

[220] Involvement of several PFC areas for the representation of a cognitive or behavioral response is explained by the recognition that such a response relies on the representation of a distributed Structured Event Complex with multiple attributes that range from motor responses to social norms (Barbey, Krueger et al. 2009).

agreement is lacking regarding automatic processing (Evans 2008). Nonetheless, there are proposals regarding the neural correlates for the systems underlying automatic and for controlled processing respectively. Not surprisingly, a shift from controlled to automatic processing implies a diminishment of prefrontal lobe activity, which is recruited for controlled processing. The automatization of a previously controlled task likely depends upon increased basal ganglia activation, inter alia (Lieberman 2007).[221] Concurring with this is the account that focuses on the necessity to answer to the capacity limitations of working memory and executive control, equally implying that improved self-regulation without further challenging these limitations should diminish the reliance on PFC activation (Deutsch and Strack 2006). The latter account also refers to research of the basal ganglia as being involved in such a shift, while acknowledging that it is still unclear how and for what specific role the basal ganglia are involved. Processes as diverse as selection, inhibition, and attention direction have been mentioned as contributing to automatic executive control (Heyder, Suchan et al. 2004).[222] In sum, in terms of neural correlates, the kludge formation might be taken to consist largely of a shift from PFC activation to basal ganglia involvement, the latter being in need of further specification depending on the task at hand.

Our fourth kludge characteristic referred to the inter-individual variation visible in the kludge formation or in its final state. Such variation can depend from the large variation that is found between individuals in working memory, which will also affect when and how a shift in their modes processing occurs (Feldman Barrett, Tugade et al. 2004). Furthermore, as much as there are several forms of self-regulation or strategies that can support a shift in processing, these will add to the large variation between individuals or even between tasks. This variation is not only observable in task performances properties, but is also associated with differences in the mechanisms responsible for these performances. Indeed, Stanovich even rejects the notion of a general and task-independent automatic system but instead prefers to refer to TASS or 'The Autonomous Set of Systems', each of which can function as a distinct, autonomous system and can be triggered separately (Stanovich 2009). This emphasis on inter-

---

[221] Lieberman derives from his review of the literature on dual process theories the following neural correlates: automatic or reflexive processing is carried out with recruitment of amygdala, basal ganglia, ventromedial prefrontal cortex (VMPFC), lateral temporal cortex (LTC), and dorsal anterior cingulate cortex (dACC) and controlled or reflective processing recruits primarily lateral prefrontal cortex (LPFC), medial prefrontal cortex (MPFC), lateral parietal cortex (LPAC), medial parietal cortex (MPAC), medial temporal lobe (MTL), and rostral anterior cingulate cortex (rACC) (Lieberman 2007).

[222] A comparative account of the basal ganglia suggests that they should be considered a 'specialized device for the solution of selection problems' and can take over selection processes from the cortex, with which it is tightly connected (Redgrave, Prescott et al. 1999). Selection being an important and initial step of information processing, this account may at least explain part of the effect of automatization and expertise.

individual variability concurs with our earlier observation of different strategies that might be involved in a shift of processing, which will recruit more or less different neural systems as well.

Our fifth kludge characteristic refers to its employing pre-existing components, that might also be involved in other mechanisms. Similar to the fact that developmental learning was found to be largely a matter of modularization of processes that were already given (Karmiloff-Smith 1992), kludge formation in the present context equally depends on the 'neural re-use' that is prevalent in cognition and the brain (Anderson 2010). As noted earlier, particularly in section II.3.1.1, the assumption that two different processes can be distinguished does not imply that the two are strictly separate, nor that they rely on completely different systems. Indeed, when two different processes relying on different memory systems are apparently responsible for psychological and neuropsychological results, it is still useful to not strictly separate them but to recognize possible interactions between the two when modelling such results (Smith and DeCoster 2000). Even automatized reasoning tasks are shown to possibly invoke both processes, contrary to the belief that such forms of rationality can only be subserved by controlled processing (Stenning and Lambalgen 2008). More explicit in denying strict separation of the two types of processing is the uni-model account that assumes that a shift in processing merely involves the differential activation of a single model in response to specific parameters of the task at hand (Kruglanski and Orehek 2007) – suggesting a comprehensive explanatory mechanism which responds in a slightly modified manner to specific tasks, comparable to shifting gears or employing different components. Similar is the distinction of not two but four processes that interact differentially in task performances, in response to various task features (Sherman, Gawronski et al. 2008). In sum, dual-process theories do not assume that the kludge formation involved in shifting between processes depends upon the establishment or deployment of a novel mechanism.

Are kludges in this context involved in further dynamic trajectories, like being partly responsible for subsequent development or learning? Or are such kludges highly specific and only activated in very limited situations? Discussion of this sixth characteristic can be relatively short: insofar as kludges function as means to decrease processing demands, they allow subjects to perform increasingly complex and novel tasks in which such kludges are integrated. Indeed, automatization of a task via chunking helps to reduce the relational complexity of the relevant representations, allowing an agent to subsequently engage in even more complex tasks (Halford, Wilson et al. 1998). In the previous chapter and in the next part where the hierarchical structure involved in cognition and behavior is at stake, we argue that the performance

of increasingly complex tasks is dependent upon an agent's ability to form kludges, as when a shift to automatic processing of a particular task component occurs and allows him to shift his attention to another task component.

The final and seventh kludge characteristic pertains to the integration of environmental information. Not surprisingly, as with every developmental and learning process, environmental information is comprised in the eventual kludge. Indeed, evolution prefers open mechanisms precisely because they allow such including of environmental information, as can be observed once the imprinting mechanisms of chicks have integrated detailed information about their caregiver (Wimsatt 1986).[223] As kludge formation often involves an automatized informational process in response to specific environmental stimuli, it is plausible to assume that such information determines the kludge to a large extent. Interestingly, confirmation of this comes from a paradigm for targeted adjustment of an agent's kludge, which consists of his learning to automatically associate the specific stimulus with another response direction (Hofmann, Friese et al. 2008). Another self-regulating strategy influences kludge activation by using representations of a future task situation in which an intention should be implemented (Gollwitzer and Sheeran 2006), usually including both perceptual and linguistic information.

In the next chapter, we will more specifically consider how humans are particularly apt at kludge formation with the integration of environmental information, tools and language. In a certain sense it will complement our preceding arguments concerning strategies that aim to modify the representations involved in the relevant processes without calling upon other resources.

---

[223] It has even been argued that we should recognize that genes, too, contain information that is partly adapted to the environment and history of the organism (Collier 1998).

## 4  THE BRAIN AS A MECHANISM CAPABLE OF KLUDGE FORMATION AND OPEN TO EXTERNAL INFORMATION

We began this part by noting that development and learning allow a baby to expand its vocal repertory from mere babbling without clear rhythm and distinct melody to a wide range of vocal expressions. At the other end of the spectrum, an expert singer can learn complete opera roles that require extensive singing and acting performance simultaneously. According to the neuroconstructivist account of development and learning, the mastery of such vocal skills corresponds with a process of modularization. This process is associated with both the proceduralization or automatization of these skills and the explicitation or increasing capability to articulate the relevant representations, allowing an expert to adjust and correct his cognition and behavior (Karmiloff-Smith 1992 ; Mareschal, Johnson et al. 2007).[224] Related to these processes are – usually implicit - processes that result in shifts between types of information processing, aiming to mitigate the consequences of capacity limitations in the brain, like the strategies of segmentation or chunking of information (Halford, Wilson et al. 1998).

Such shifts in processing or skill mastery can be illustrated with the example that we used in this part as well: singing. For we discussed the fact that expertise in singing opera roles can result not only in learning different roles – like a Don Giovanni, a Wotan, and a Saint François – but also in learning to interpret these roles according to the requirements of a specific interpretation of the opera character. This can lead to a situation where an expert singer has interpreted a role in conflicting ways, requiring him to inhibit a stereotypical rendition of a virile Don Giovanni or serene Saint François in order to give way to an alternative interpretation. Interestingly, although much of our human behavior and cognition appears to be determined under circumstances by either automatic or controlled processing, we still have the capability of bringing about a shift from controlled to automatic processing. For the automatization of controlled processing we have several strategies available, such as those mentioned in the previous

---

[224] As discussed in section II.3.1.5, our account differs from another account of skill acquisition that sharply distinguishes between expert intuition and the analytic reasoning that agents rely on before reaching that expert level (Dreyfus and Dreyfus 1986 ; Dreyfus 2004). In contrast, our account of kludge formation aims to explain how it is possible that experts are capable of fighting the capacity limitations associated with controlled processing in such a way that they remain capable of articulating or making explicit what they do. We don't want to deny that novices and experts share their reliance upon implicit knowledge, while intermediate experts will more often invoke explicit knowledge. However, it is quite possible for implicit knowledge to be made explicit, as when a neural network can offer a representation of its own specific state (Cleeremans 1997 ; Cleeremans, Timmermans et al. 2007). What we do deny, therefore, is that experts rely on a critically distinct process, different from those that novices use. Indeed, above we refer to an updated account of Simon's concept of chunking, that can explain why an expert may need to learn how to articulate his chunked and implicit form of information processing (Gobet and Chassy 2009).

section. These have observable results even in a situation where common controlled processing will not be viable due to pressure or the availability of limited cognitive resources. Apart from strategies aimed at a reduction of the information load of the upcoming process by segmenting or chunking the relevant information, an agent can also engage in self-regulation by employing explicit representations of the future situation or future response in order to prepare for a controlled yet quasi-automatic performance of the desirable behavior.

All such learning processes can be considered as dependent upon the formation of a kludge modifying the mechanism that can be held responsible for a singer's performance. That is, in all such cases the explanatory mechanism's activity leads to observable differences in its behavior, even though it may have remained largely the same as before. It may be that a kludge has formed such that keeping tone no longer requires ongoing attention, but is automatized. Or a kludge has been added to the vocal expression mechanism, enabling the singer now to reliably coordinate his singing with keeping rhythm. The kludge formations that we have observed so far appear as internal modifications, requiring no apparent resources from outside the agent. However, this may already be an oversimplification. In the next chapter we will discuss two types of kludge formation that obviously involve external resources.

After all, kludge formation that obtains during singing practice is not just an internal affair comparable to repeating a spontaneous behavior many times. Although our first kludge characteristic focuses on its functional properties, these may be shaped or sculpted under the influence of external or environmental information – which refers to our seventh kludge characteristic. Obviously, with this combination of characteristics, we can expect kludges to emerge in practically all domains of cognitive processing. An example that demonstrates the interaction between developing functional properties and environmental information is a baby's spontaneous babbling or crying which gradually gives way to more melodic vocal expressions. Contrary to the assumption that this process depends just upon repeating prespecified behavior, culturally specific external information about tonal mode turns out to play a role here from the moment that a baby starts to cry and increasingly so when specific scales are used for music and in singing.[225] On top of these culturally specific influences, there are many other

---

[225] Newborn infants cry differently in France and in Germany, reflecting the specific inflections of their respective mother tongues (Mampe, Friederici et al. 2009). As is well known, during their youth, children's capacities for the general recognition and production of specific vowels and consonants are increasingly constrained under the influence of exposure to their mother tongue(s). Similarly, the specific rhythms and scales prevalent in the environment will shape the space of a person's vocal expressions, even if he may later be able to expand or transgress that shape with effort. Indeed, young children have better memory of absolute pitch, probably because it allows them to better recognize the voice of their caregivers (Trehub and Hannon 2006). But even after early youth, the exposure to a pitch language influences absolute pitch recognition, explaining why Japanese children are better than Canadian children at memorizing pitch (Trehub, Schellenberg et al. 2008).

resources that an opera singer must learn to integrate in his performance. Consider, for example, the abundant use of external tools and information, like a music score, pictures, stage-properties and the like. Interestingly, performance is not always made more complicated by these external additives. On the contrary, performing a role can be facilitated by their use, as when Don Giovanni's or Saint François' particular singing lines become associated with the singer's successive and specific manipulations of a sword or cross, which are generally easier to remember. Even amateurs can experience that upon seeing a long forgotten score their voice automatically prepares for its first, high and dissonant, note.

It is to this phenomenon of integrating environmental information in the formation of a kludge that we now turn. Although it has already been mentioned earlier, it is of such importance that we will not close this part before treating it separately. It will confirm our earlier observation that our behavior and cognition can be considered to be the result of extremely 'open programs' (Mayr 1974) capable of integrating such information. Indeed, this integration can occur at such an early stage of development or learning that its result in turn is employed in subsequent behavioral or cognitive mechanisms. At first sight, such deeply and generatively entrenched kludges often look like innate mechanisms, although in fact they are acquired (Wimsatt 1986). In what follows we will discuss a few accounts of kludge formation with the involvement of external information. It will conclude our preparation for Part III, that will put forward a combined philosophical and cognitive scientific approach to a hierarchy of intentions.

## 4.1  Symbols, simulators and the malleability and stability of cognitive processing

With the two previous discussions on kludge formation, focusing on child development and learning, and on dual-process theories respectively, the impression may be that kludge formation comes naturally and only employs natural resources that are available without an important role for socio-cultural or environmental information. In order to ward off this impression, the present section will discuss theoretical arguments and evidence suggesting that humans are capable of developing 'simulators' such that these tie together a variety of content features employed in different conditions and for different purposes. This concurs with the general observation, which we referred to several times above, that evolving and developing mechanisms will usually benefit from the capability of integrating such environmental information (Wimsatt 1986 ; Wimsatt 2001). In what follows, we will observe that not only environmental information, but also internal information that is stored in a distributed distributed

way, can become integrated similarly in such mechanisms. This bestows upon such dynamic mechanisms several highly relevant properties.

Obviously, many dynamic mechanisms are capable of responding to environmental information on a momentary basis, allowing them to respond to environmental changes. However, in this chapter we are also interested in the integration of such information in a more enduring and stable form in mechanisms. If that happens, it will allow the further development of such a mechanism to build upon the integrated information, which will as a consequence become ever more entrenched into it. This malleability of the mechanism does not undermine its stability but instead enhances this, as the integrated information itself has contributed more directly to the mechanism's functional properties. As a consequence, the mechanism will generally respond more reliably and faster to related environmental information. The question that poses itself, however, is whether this development robs the mechanism of some of its adaptivity or flexibility, being less capable of responding to unexpected and novel environmental features. We will argue that a mechanism's adaptivity combined with its stability will depend upon the hierarchical structure of the information it has integrated. If it is capable of preserving a necessary hierarchy of this information, it may be able to process continuously novel environmental features with a largely unchanging and stable mechanism.[226] The next sections will be devoted to the discussion of kludges that are constituted by a cognitive mechanism's interacting in an enduring and stable manner with specific environmental and internal information. Our discussion of this capability will first discuss a theoretical account of cognition that assigns an important role to cognitive simulation. Subsequently, we will discuss the concept of extended cognition, which implies that cognitive processes do not occur solely inside the skull but extend into the world, integrating in a tight manner not just environmental information but even objects and technologies.[227]

Although several simulation theories are on offer (Goldman 2006 ; Hesslow 2002 ; Zwaan 2008), we will focus on Barsalou's account, as it develops its argument such that it simultaneously allows a central role to symbols and language in human

---

[226] The prevalence of theories that ascribe hierarchical structures to cognitive processes and representations is largely due to the assumption that such a structure offers several benefits, such as stability and speed in processing, multiple forms of access, and many combinatorial options (Cohen 2000)

[227] The debate concerning the 'extended mind' has meanwhile been reformulated as 'the Hypothesis of Extended Cognition' (Clark 2011), which also better suits our vocabulary.

[228] Zwaan focuses on language processing and offers a simulation account of language comprehension. Mental simulations are proposed for explaining the structures that can be noted in our comprehension of events or situations (Zwaan 2008). However, this account implies that sensorimotor activations in particular are involved in language comprehension, and is less interested in how language or symbols are stored in the brain than Barsalou's account is (Barsalou 1999c ; Zwaan 2009).

cognitive processing.[228] We will start our discussion with this latter issue. Accounting both for abundant evidence of many forms of interactions or interferences between action, perception and speech that is visible in subjects' performances and for the development of apparently abstract concepts in human speech, Barsalou developed a theory of perceptual symbol systems (Barsalou 1999c). Though this may not be the only proposal for explaining those forms of interactions or for grounding concepts in multimodal bodily experience, it has two interesting aspects that merit specific attention.[229] First, the theory carefully argues against the suggestion that sensori-motor experiences are stored holistically, in which case the correspondingly grounded concepts would presumably represent an object, event or action equally holistically. Instead, the theory of perceptual symbol systems defends the view that memories are stored in a distributed manner across the brain - not in a holistic manner but divided in many components or features. Indeed, dependent upon the subject's attention in a given situation on specific features of an object or event or action while neglecting other features, the composite yet structured symbol that is stored will comprise components different from those that would have been stored in another situation (Barsalou 1999c ; Barsalou 2009). Such a modal symbol is stored in long term memory in a distributed manner, where connections between perceptual, motor, cognitive and affective features of an object, event or action is formed by means of numerous associations (Barsalou 2008 ; Goldstone and Barsalou 1998). Based upon the strength of the associations between features, a result of their regular co-occurrence or their being included in the subject's focus of attention, these associated features are likely to be co-activated in subsequent situations.[230] This brings us to the second aspect of this proposal.

The account denies that the retrieval or activation of memory of an object, event or action involves the reactivation of a holistic and faithful representation of it, as though it were stored comprehensively in memory. Rather, a memory activation requires the re-activation of a complex of several different features, stored at different locations in the brain and associated more or less strongly with each other. Barsalou calls this process

---

[229] An earlier proposal for symbol grounding which also aimed to account for the presence of abstract and concrete symbols was presented in (Harnad 1990). Its original aim was more modest, as it did not seek to account for the interaction between cognitive functions in which symbols may play a role, nor for the top-down influence of symbols on perception, for example.

[230] However, they need not be co-activated all the time, also allowing for the development of novel, amodal symbols. Indeed, even though abstract thought may have perceptual origins, it does not preclude the creation of abstract contents that comply to new rules (Goldstone and Barsalou 1998). Besides, abstract content may have an origin not just in perceptual and motor states, but also from internal states or introspection (Barsalou 2008). In any case, it would be problematic to derive all possible conceptual contents and structures directly from modal symbol systems, without any role for amodal and abstract symbols (Dove 2009).

simulation, about which he writes: "[a]ccumulating evidence suggests that simulation constitutes a central form of computation throughout diverse forms of cognition, where simulation is the re-enactment of perceptual, motor and introspective states acquired during experience with the world, body and mind" (Barsalou 2009 1281; cf. Barsalou 2008)).[231] Such re-enactment occurs in a multi-modal fashion, drawing together again many features that were stored in a distributed way following many experiences.[232]

The brain facilitates such comprehensive re-enactment with the 'simulators' that integrate the relevant contents, emerging as a result of repeated experiences. Subjects will develop such simulators in great numbers, depending on their attending to actions, introspections, objects, events, situations, etcetera. The activation of such a simulator will occur upon perceiving a relevant situation or upon using a relevant word or concept. It can involve not only the re-enactment of modal states but also of conceptualizations or motor states (Barsalou 2009). This is the reason why simulation appears to be 'a central form of computation'.[233]

Importantly, such an activation or re-enactment of a simulator is always a dynamical process and will yield a visual, motor, affective or conceptual simulation that is partly dependent upon the specific situation the subject is in. As a result, a simulator activation in a given situation will yield simulations that differ between subjects, and may also vary from situation to situation for a single subject, depending partly on the strength of associations that have become integrated in the simulation (Barsalou 2002 ; Barsalou 2009). However, simulators are not stored or activated randomly. Interestingly, they do have a structure that renders them both flexible and with some stability: a hierarchical structure. It consists of a superordinate 'frame' level,

---

[231] Simulation gained much interest after the discovery of mirror neurons, which support the activation of highly similar neural networks under different conditions, for instance upon the observation, imagination, imitation, verbalization and performance of actions (Grèzes and Decety 2001 ; Hesslow 2012). It is hypothesized that mirror neurons are also implicated in the evolutionary development of language, gestures forming an evolutionary precursor to language. See e.g. (Arbib 2003a ; Arbib 2005 ; Jeannerod 2008 ; Rizzolatti and Arbib 1998). Barsalou recognizes this potential relevance of mirror neurons for simulation theories, yet raises the question why non-human primates show such different abilities even though they have mirror neurons at their disposal, too (Barsalou 2008).

[232] Partly because of the involvement of mirror neurons, simulation gets its experiential, multi-modal nature, making it a richer source of understanding other subjects than theorizing, according to Gallese & Goldman's account of intersubjective understanding (Gallese and Goldman 1998). Nonetheless, most cases of mindreading are likely to depend upon a hybrid of both mirroring and theorizing, according to Goldman (Goldman 2006).

[233] Concurring with this observation, Hesslow defends a "'simulation' theory of cognitive function", arguing that a wide range of cognitive processes in fact converge on their being a form of simulation. Most cognitive processes are one of three forms of simulation: simulation of action, simulation of perception, or anticipation by way of simulation of action consequences (Hesslow 2002,242).

comparable to the category class, and a subordinate level, which includes contents referring to specific members of that category – such as specific cars that all fall under the frame 'car' (Barsalou 1999c, § 2.4).[234] With this structure and with the possibility of combining the many components in indefinitely many ways, these simulators can also account for the productivity that characterizes language. What is more, simulators are also involved in the development of abstract concepts (Barsalou 1999c ; Barsalou 2008 ; Goldstone and Barsalou 1998).[235]

A final aspect of this account that merits our attention is the fact that as much as simulators may give rise to linguistic concepts, concepts in turn can activate and control the simulations that rely on such simulators (Barsalou 1999c). Instead of assuming a uni-directional influence of modal systems on conceptualizations, the account recognizes a bi-directional interaction in which rich experiential simulations can also result from a subject's attending to a concept (Barsalou 2009 ; Barsalou, Cohen et al. 2005).[236] Indeed, this concurs with Barsalou's general assumption that language comprehension is not so much a matter of activating an archival memory but is better considered as preparing agents for situated action, "or at least to create the experience of situated action" (Barsalou 1999a 75).[237]

Experience and expertise play a crucial role in this account. Obviously, an agent's simulation of a situated experience and his simulation of potentially adequate actions will be richer if he can rely on a rich history of many relevant experiences

---

[234] The multilevel, hierarchical structure of a simulator is comparable to the account of chunks in (Gobet and Simon 1996), where a chunk consists of a template with several free slots that allow situation-dependent specification. Both theories refer to hierarchical structures and thus aim to account for easy and fast recognition of complex content while avoiding a holistic account of the activated, memorized content.

[235] While referring to this structure and the consequent productivity and recursivity of language, Goldstone & Barsalou argue that instead of strictly distinguishing between perceptual and conceptual representations, and between perceptual and abstract contents, we should conceive these as continua (Goldstone and Barsalou 1998). First empirical neurophysiological evidence in support of this theoretical continuity came from TMS experiments, showing that subjects activate motor areas for processing both concrete and abstract concepts (Glenberg, Sato et al. 2008). Meanwhile, other lines of research have supported this evidence that sensorimotor areas are recruited for processing abstract concepts (Ghio and Tettamanti 2010 ; Lacey, Stilla et al. 2012 ; Pecher, Boot et al. 2011 ; Pulvermüller 2012). Barsalou's theory is specific for its explanation of the combinatorial productivity of language and its offering a richer account of abstract concepts than merely considering these as impoverished concepts for concrete objects (Barsalou, Simmons et al. 2003).

[236] In a different context, Goldstone notes that the handling of abstract mathematical operations has an impact upon subsequent perceptions made by a subject. It once more confirms the widespread phenomenon of neural re-use (Anderson 2010) of evolutionary early – perceptual - systems for later cognitive – mathematical – tasks (Goldstone, Landy et al. 2010). The simulators and simulations that are discussed in the present section are another example of such re-use.

[237] Obviously, the idea that the brains of organisms store information in order to better prepare for future actions is not specific for this account. Such a notion of memory as a source of information for prediction is made more explicit in (Bar 2009) – which account also assigns an even larger role for top-down information or analogy seeking in the processing of perceptual information.

(Barsalou 1999a).[238] Moreover, given enough experience, a simulator for a particular skill can become so entrenched that it may be activated automatically in a situation that appears to be relevant, without requiring conscious effort.[239] Such simulator activation or re-enactment will also enable the subject to predict or anticipate how the situation will develop, as the simulator involves components that are not yet visible in the situation but that have become associated with previously experienced, similar situations: future actions, necessary instruments, likely personal feelings, etcetera (Barsalou 2009).[240]Finally, and crucially, expertise will play out to the extent that experts will be more capable of exerting selective attention to relevant as opposed to irrelevant features of a situation or object, which influences subsequent processing of these (Barsalou 1999c).[241] As we already mentioned, a simulator is never a faithful re-enactment of a given object, action, situation and so on, but can draw together

---

[238] This account clearly differs from Sperber's modularity thesis that ascribes 'massive modularity' to the human brain, claiming that it develops modules not just for perceptual functions, but also for many quite specific cognitive processes and  contents. According to this thesis, even 'micro-modules' exist that have as their domain just a single concept (Sperber 1996). Applying the notion of modules so widely appears problematic, as it must subsequently explain why perceptual and conceptual processes interact or respond to similar features of a category, or why category mistakes occur – which is better accounted for by the simulators described in (Barsalou 1999c). Although there is no room here to expand on this, we also believe that the recombinatorial productivity that is visible in human concept use is easier to explain in terms of simulators than in terms of micro-modules.

[239] As the simulators emerge as a result of recurrent associations, statistical processes in the brain are involved in several aspects of symbol and language processing, contributing to an ever more refined structure of those simulators (Barsalou 2008). Notwithstanding these statistical processes, there is a significant top-down contribution of specific attention or concept use that influences the activation of one or another simulator.

[240] Similar to this account of simulators is the hypothesis that human concepts emerge from distributed activation patterns in sensory and (primary) motor areas following upon experience. A concept then functions as a 'cog' – a term that is again similar to our notion of a 'kludge' – in our brain, providing "general structuring for sensory-motor observation, action, and simulation" (Lakoff 2006 161). However, the authors take the embodied nature of concepts much stricter. For example, they assume that these cogs provide image-schemas like containment schemas, source-path-goal schemas, force dynamics schemas, orientation schemas and others grounded in bodily experiences. Moreover, the authors suggest quite a tight relation – including some isomorphism - between the role of a concept and the functional structures of the associated brain networks: "the inferential structure of concepts is a consequence of the network structure of the brain and its organization in terms of functional clusters" (Gallese and Lakoff 2005 468, in italics in original). They go much further than Barsalou in this, and don't seem to recognize a bi-directional influence between concepts and the simulations they may provoke. Finally, they attach a certain rigidity to the functional structure of cogs, which is not the case for Barsalou's simulators – which have characteristic 'open-endedness' and allow for multiple reconfigurations (Barsalou, Kyle Simmons et al. 2003) -, or for our kludges.

[241] Treisman correctly notes that attention may not be a 'unitary process', as it both precedes and contributes to our perception of a coherent scene, and as it may rather be the outcome of a competition between different objects or features for our attention (Treisman 1998). We may expect that expertise can have a specific role in this, as it supports the outcome of such a competition precisely due to biases that are a result of previously accumulated perceptual and cognitive experiences.

any possible configuration of their components – thanks in part to the flexibility of our attention (Barsalou, Simmons et al. 2003). Moreover, and this may be peculiar to humans, such attention may be directed jointly by two subjects, thus capable of coordinating their handling of a shared situation or object (Barsalou 2005). The latter case of joint attention will benefit when both subjects share largely the same kind of expertise, facilitating their attending to the same relevant versus irrelevant features.[242] Now that we have pointed out that expertise is reflected in the development of simulators and the subsequent ease with which simulations of objects, situations, experiences or actions emerge, let us close this section by drawing another comparison to our notion of kludge.

## 4.2  Simulators and the kludge characteristics

Do simulators according to Barsalou's account have properties that make them similar to kludges, in that they affect a mechanism, or activate it in a specific way such that certain cognitive or behavioral responses can be explained? Does this account acknowledge the impact of learning and experience such that its description of a simulator concurs with the seven characteristics that we attached to the notion of kludge?

The first and foremost characteristic of a kludge was that its functional properties, rather than other properties, demonstrated its emergence. At first sight, it may seem that the theory of perceptual symbol systems is more interested in its neural properties than functional ones, as Barsalou writes: "the basic definition of perceptual symbols resides at the neural level." However, he continues with a clarification which shows why it is precisely from the functional property of conscious availability that we can infer a very general neural property of perceptual symbols: "*unconscious* neural representations – *not conscious* mental images – constitute the core content of perceptual symbols" (Barsalou 1999c 583, italics added). Indeed, even though his account ascribes a central role to certain -  multimodal – processes that underlie simulators and simulations, it primarily aims to account for many functional properties of human cognition and behavior. For its main theoretical purport is to contest the division between cognitive functions like perception, cognition and action. It does so by criticizing an amodal account of distinct concepts or categories and by building a theory about how subjects' concept use depends upon multimodal simulators that enable them to prepare for

---

[242] Obviously, however, in case attention needs to be drawn to a previously unobserved feature, expertise may have an impeding effect. Nonetheless, given that expertise results in more refined and structured simulators, Barsalou's account suggests that even in such cases expertise may support the discovery and interpretation of such novel features.

situated action (Barsalou 2002 ; Barsalou 1999a). Not surprisingly, therefore, such preparation for situated action is evident from functional properties produced by the simulators, primarily in modified behavioral responses resulting from the development of such simulators. Similarly, they appear to affect sensory perception, as research has shown. For lesion studies suggest the involvement of particular neural areas in simulator development, but they do so by providing evidence that consequences of lesions are observable in patients' affected behavior or cognition, thus supporting the focus on functional properties of simulators (Barsalou 2009 ; Barsalou 2008).

Our second kludge characteristic implied that we cannot derive directly from a kludge's functional properties what algorithmic theory can describe its operation. Now Barsalou appears to have committed himself to a particular algorithmic theory - in Marr's sense (Marr 1982) - in which simulation plays a central role in different forms of information processing ranging from perception, speech, and action to introspection, as we learnt in the previous section (Barsalou 2008 ; Barsalou 2009). However, the specific form of the process involved in a particular simulation can be highly varied. A simulator refers to a 'distributed multi-modal system' that integrates an increasing amount of information as diverse as properties, relations, events and mental states that are related to a particular category, like cat or bicycle (Barsalou 2009). Such simulators are developed for: "any component of experience (or configuration of components) processed repeatedly by attention" (Barsalou, Simmons et al. 2003 89), including not just attended objects or a situation external to the subject but also those that are present during introspection. As a result of these processes underlying a simulator, it is indeed unlikely that we can derive from a particular simulator's functional properties what representations have been employed nor how these have been use. Think of an expert singer who can prepare his role by either drawing upon a specific experienced situation, or upon an idiosyncratic configuration of some features of a typical situation, or upon his repeated imagination of a certain opera scene.

As for the neural implementation of the 'distributed multi-modal systems' that underlie simulators, it is to be expected that this third characteristic cannot be very specific.[243] Nonetheless, a general proposal has been made in (Simmons and Barsalou 2003), where the authors distinguish between at least two different components of the development of simulators and their neural implementations. To begin with, given that simulators are developed from situated experiences, sensory-motor

---

[243] Barsalou acknowledges how its distributed nature also plays out in the equally distributed, underlying neural systems: "a situated conceptualization is a multi-modal simulation of a multi-component situation, with each modal component simulated in the respective neural system" (Barsalou 2009 1283).

systems are intrinsically involved. Experimental neuroimaging and lesion studies do indeed demonstrate that specific auditory, visual, and motor systems are activated when subjects process words related to specific categories or concepts (Simmons and Barsalou 2003).[244] But a simulator is not just the re-enactment of an actually experienced sensori-motor activity, as a simulator typically also integrates information stored in memory and derived from other experiences or simulators that share a feature property with the activated simulator. Neural association areas are suggested as the neural implementation for such sharing of features among simulators, comparable to the hidden layers in a connectionist feed-forward network (Barsalou 2003). The idea behind this hypothesis is that conjunctive neurons reactivate the components of other distributed networks, underlying other simulators, that are related to a currently activated simulator (Simmons and Barsalou 2003).[245] Other neural areas can be recruited as well to support a simulator, for example when introspective states are employed for the simulation of particular features or contents (Barsalou 2003) or when imagination creates novel and unrealistic simulations (Barsalou 1999c). In sum, neural implementations of simulators differ largely and may even differ from time to time when a specific simulator is activated, depending again on the selective attention or context in which it is activated.

That brings us to the fourth kludge characteristic, which hardly needs specific attention. As a simulator is continuously developing and as context and attention determine the specific informational contents that are being activated in a given situation, its variability and flexibility is large. This does not imply that each individual has a completely random set of simulators with idiosyncratic informational properties, fortunately. Given the important role of modal and embodied processing and the influence of statistical processing of environmental information for the development of simulators, many simulators and their properties will largely remain stable for a subject and will to a great extent be similar among subjects (Barsalou 1999c ; Barsalou, Simmons et al. 2003).[246] Indeed, two subjects need not have developed identical simulators and can still produce strikingly similar simulations of a concept in a given

---

[244] The involvement of sensori-motor systems in language processing meets with relatively broad agreement. See e.g. (Beilock 2009 ; Deacon 2006 ; Jirak, Menz et al. 2010 ; Pulvermüller and Fadiga 2010 ; Tremblay and Small 2011).

[245] The authors refer to Damasio's idea of 'convergence zones', which equally aims to explain cognitive processing of language and meaning as the product of distributed networks, with an important role for various types of convergence zones, responsible for the abundance of associations observable in language and meaning processing (Damasio 1989). A recent review on embodiment and semantic representation concludes that in addition to the crucial role of sensori-motor areas, different theories are 'converging on convergence zones'. The reviewers point out that on the basis of the literature it is plausible that processing of abstract concepts does not necessarily engage primary sensori-motor areas  - as strong embodiment theories would have it (Meteyard, Cuadrado et al. 2012).

situation or text, for example when this text or situation offers enough constraints on their simulations (Barsalou 1999c ; Barsalou 2009). Obviously, the richer the amount of experience and the simulators developed by subjects are, the more likely it is that they can produce such comparable simulations.

From this last observation we can immediately derive that a simulator is 'cobbled together', as our fifth kludge characteristic states. We asserted repeatedly that simulations are not to be understood as the activation of stored holistic memories of objects, events or actions. Instead, the simulator involved stores many different yet related perceptual symbols, recruiting sensori-motor representations among others.[247] These symbols are stored in hierarchically structured networks, with frames at upper levels and further lower level contents. The development and activation of such a structured simulator is dependent upon a subject's attention, the use of concepts and their meaning, and so on (Barsalou 1999c ; Barsalou 2009). Critically, having stored many of such simulators, humans can do more than just reactivate them to simulate previously experienced objects, events and actions. On top of this, they are capable of developing novel, abstract concepts and of simulating nonpresent situations with each other, which may be a critical component of human evolution (Barsalou 1999b). In sum, a simulator can be considered as a kludge also in the sense that it is composed as a distributed yet structured network consisting of components that are involved in many other cognitive processes as well. Notwithstanding this composite nature, each simulator has a relatively stable – hierarchical – structure, which merits its being called a kludge.

Being cobbled together, does a simulator in turn play a role in subsequent developments or experiences and become increasingly entrenched, as our sixth kludge characeristic would predict? Barsalou argues that this is indeed the case and reviews evidence that shows how perception, imagination, speech and other processes are facilitated and accelerated by the presence of relevant simulators. Based upon such

---

[246] Moreover, there is behavioral and imaging evidence that in a general sense a conceptual system and even its implementation are to some extent shared between species like monkeys and humans (Barsalou 2005). A more elaborate theory of the evolutionary development of human language from precursory systems shared with animals is developed in (Arbib 2011). That account assigns an important role to mirror neurons, which appear to be recruited both in sensori-motor and in other association systems. Mirror neurons are also assumed to underlie the activation of motor systems when subjects are processing sentences with abstract conceptual contents (Glenberg, Sato et al. 2008). See footnote 231 above on the relation of Barsalou's account to mirror neuron research.

[247] There are several lines of evidence for such an embodied account of symbol and language processing, for example when subjects recruit sensori-motor areas in fMRI experiments of language processing, or when patients suffering from lesioned sensori-motor areas have difficulty in specific language tasks. Such findings suggest that sensori-motor activations are not just epiphenomenal, but also functional in language processing (Barsalou 2008). Evidence also suggests a bi-directional influence between sensori-motor activations and language processing (Pulvermüller 2012 ; Pulvermüller, Hauk et al. 2005 ; Zwaan 2009).

simulators, a subject will implicitly anticipate and attend to specific information when perceiving an occluded scene or half-understood sentence, for example, which further expands these simulators and strengthens their role in future situations (Barsalou 1999c). Concurring with these arguments, one comment adds that knowledge structures like frames and scripts – which will figure together with schemas in our next part – are related to such simulators and similarly contribute to language comprehension and other functions in the form of background knowledge (Zwaan, Stanfield et al. 1999).[248] Other support for Barsalou's observation concerning entrenchment comes from research on implicit memory, which has shown how memorized items subsequently facilitate the processing of related items by several cognitive functions like perception and the imagination of the past and future (Schacter and Addis 2007c). Entrenchment of particular simulators can even be observed in religious beliefs and rituals, which often employ contents that are grounded in modality-specific brain systems. By doing so, religious practice contributes to the social dissemination and recurring activation of these simulators (Barsalou, Barbey et al. 2005). Interacting cognitive and socio-cultural processes are thus responsible for the generative entrenchment of a collection of simulators that grows with an agent's expertise.[249]

Finally, for our seventh kludge characteristic we must consider whether environmental information is involved in the development and functioning of a simulator. Although simulators are grounded in perceptual symbol systems and differ in that respect from amodal accounts of symbols or concepts, we have meanwhile elucidated that this account emphasize the embodied nature of symbols and concepts as well as their situated nature. Environmental information does not just influence the development of simulators, often a simulator is activated in response to a situation (Barsalou 1999c). Experts, in particular, can be recognized by the fact that certain situations will automatically re-enact similar situations, with the activation of specific simulations and consequently facilitation of certain responses over others (Barsalou 2009). This results in the subject's situated experience of 'being there', in stark contrast to a subject who can only try to apply amodal concepts to objects, events and actions that are different from the ones from which they were originally abstracted

---

[248] As mentioned in a footnote 257 the evolution of language is considered to be a rather continuous process, with several species sharing many different language processing components with humans. Reviewing evidence about the language faculty in different species, the authors of (Hauser, Chomsky et al. 2002) suggest that it is communication in a very broad sense that drove this evolution, drawing upon communicative dispositions that differed from species to species.

[249] Cultural evolution is in many ways dependent upon the generative entrenchment of specific contents and practices in the behavior and cognition of individual subjects. Humans in particular use many 'scaffolding' strategies that contribute to this process (Wimsatt and Griesemer 2007).

(Barsalou 2002). Finally, culturally specific information plays an important role in the development and activation of simulators, underlining the relevance of environmental information (Barsalou 1999c ; Barsalou, Cohen et al. 2005).

In sum, the simulators this section has been devoted to turn out to comply largely with our notion of 'kludge'. Given its wide range of application to many domains of cognitive processing, it offers strong support for our argument that the explanatory mechanisms responsible for our acting are extensively modifiable as a result of learning and experience. As a result, our space of actions is being sculpted, with some options gaining in probability to be performed, whereas other actions are less likely to come to light. Next, we will discuss a final source of support for our argument about the prominence of such modification, now focusing on our capability to develop kludges with the involvement of environmental information and even environmental objects or material structures.

## 4.3 Reaching outside the skull: how can external objects become integrated?

The first chapters of Part II were devoted to processes that appeared, at first sight at least, to be part of human natural development and experience. The phenomenon of modularization in infant learning and the development of automatized processing was not shown to rely on a specific and external source of information like language, even though it is likely that instruction does play a role in those cases. In the last section and the next, in contrast, we consider forms of kludge formation that more explicitly rely on the involvement of environmental information – and here of external objects. Simulators, which we found to play an important role in storing and employing such information, helped already to demonstrate that even culturally specific information will be integrated in cognitive mechanisms, influencing the functional properties of these simulators and the simulations depending on them (Barsalou 1999c). This is not surprising, as we already argued in chapter II.1 that we should not expect evolution to give rise to closed programs only, as these would only be beneficial for short-lived organisms with little time for learning (Mayr 1974). Organisms that live longer conversely benefit from their ability to entrench environmental information in mechanisms responsible for their behavior and cognition, with this entrenched information subsequently determining in part their development (Schank and Wimsatt 1986).

Such entrenchment of information during processes of kludge formation invites us to pause for a moment to reflect on the nature of cognitive processing. To the extent that cognitive processing involves external information, its results are always influenced by

it, obviously. Given Hebbian learning processes, these processes themselves will be modified according to the specific amounts and importance of different contents that they are offered for processing, including contents that derive from the environment: a vulnerable and quick rabbit will gain perceptual and motor expertise that is vastly different from the expertise of a human opera singer.[250] When environmental information becomes entrenched with the formation of a kludge, it thus contributes favorably to the facilitation of future processes within a specific environment in comparison to processes that are triggered by a completely novel environment.

What should be emphasized, moreover, is that the external information that becomes entrenched in cognitive mechanisms can have a profound effect on the structure of subsequent processes. In addition to associations that have been gradually and mostly implicitly shaped between specific contents – like those determining the results of automatized processing according to the dual-process theories discussed earlier – cognitive processes can follow rules or structures derived from external information. In humans, this capacity of structuring and restructuring cognitive processes according to acquired, external information is strongly developed. Language and the use of symbols play a significant role in this capacity, even if this role is different from case to case. For example, apart from facilitating agents to restructure their space of actions, symbol use can also help to inhibit automatic, reward related responses to stimuli. A demonstration of this was given in research with chimpanzees. These had difficulty in maximizing their rewards in a reverse-reinforcement contingency task in which they would receive maximum reward only when they selected the unappealing minimum, and vice versa. They failed only when they saw the attractive candies involved, but not when these were represented by abstract numbers: then the chimps were capable of selecting the lower number, obtaining maximum rewards (Boysen, Berntson et al. 1996). Symbolic representation thus helped them to override a strong action tendency and it facilitated restructuring the relation between direct aim and the indirect outcome of their action.

Compared to animals, human capacity for language and symbol learning is significantly different and larger.[251] Indeed, this capacity is so crucial to human existence because our brains have co-evolved with language use and our prefrontal cortex has overdeveloped such that humans: "are not just adapted for symbol learning but for *fail-*

---

[250] This is not to deny that there are also many innate predispositions or prespecifications at work that help determine the early learning outcomes of rabbits versus humans and their subsequent chances of gaining expertise in running versus singing. Prespecified biases play a role in the neuroconstructivist account of such expertise, seeking a middle ground between innately specified content-specific modularity and unconstrained 'tabula rasa' accounts of cognition (Karmiloff-Smith 1992 ; Mareschal, Johnson et al. 2007).

*safe symbol learning*" (Deacon 1997 415, italics in original). Meanwhile, humans have created an ecological niche strongly determined by this language capacity (van der Lecq 2012). Even though animals share some capacity for symbol learning, humans are particularly able to acquire and recompose symbolic structures as a means to prepare for novel environments and novel actions, instead of merely relying on the strength of previously exercised associations, the way animals do (Barsalou 2005). Indeed, acquiring insight in deeper formal structures facilitate the transfer of agents' expertise in pattern learning from domain to domain, even when these are different such as letters and animals. Generally, it appears that even early perceptual processes learn to adapt to acquired formal rules and process visual stimuli according to those rules. For example, eye-tracking research shows that adept readers focus on multiplication problems before focusing on addition problems in mathematical formulas. Given this impact of mastered abstract mathematical rules and structures on human perceptual processes, these processes are said to be 'rigged up' with mathematical expertise (Goldstone, Landy et al. 2010).[252]

One could consider, as we do, these phenomena as proofs of the brain's – and our human brain's in particular – potential to integrate environmental information in such a way that it is not only deeply entrenched in its mechanisms, but also generatively involved in further developments and cognitive processes. Another view is to focus less on the integration of external information, but rather on its use or employment in creating novel and 'hybrid' cognitive processes, or in 'extending' the mind. Reasoning in that vein, Clark interprets such phenomena as a result of cognition being partly dependent upon a 'symbolic environment' which offers "additional fulcrums of attention, memory and control" (Clark 2006 300).[253] This phrasing suggests, however,

---

[251] Even doves are capable of learning categories and simple categorization rules (Ashby and Ell 2001). Nonetheless, some continuity in symbol learning should not blind us for crucial differences in animal and human language capabilities. Whether it makes sense to distinguish between a broadly and narrowly conceived language capacity in order to differentiate between both capabilities requires a discussion beyond the confines of this dissertation (Hauser, Chomsky et al. 2002). What is more relevant to our purposes is the extent to which language and action processing are intertwined, which will be discussed more prominently in the next part. Interesting in that respect is the fact that Broca's area is involved in processing not just complex structure in language, but also in action (Arbib and Bonaiuto 2007 ; Hamzei, Rijntjes et al. 2003 ; Koechlin and Jubault 2006 ; Nishitani, Schurmann et al. 2005 ; Willems and Hagoort 2007).

[252] Conversely, bodily gestures appear to scaffold our acquisition of mathematical cognitive operations, facilitating learning and influencing brain connectivity correspondingly. At later stages, gestures still play a modulating role during mathematical performances. De Cruz considers this to be supportive evidence for the extended mind hypothesis (De Cruz 2008). Dehaene, in his work on mathematical representations and reasoning, does not draw such consequences even though he acknowledges the transformative impact of linguistic – discrete - number representation on the (encultured) brain (Dehaene 2001 ; Dehaene, Spelke et al. 1999).

that the environment remains external to those cognitive functions, while we have been emphasizing the integration of external information in cognitive mechanisms in such a way that it not just modulates but actually modifies those mechanisms and thus the cognitive processes they perform.[254] For Clark, however, the relative externality of this information is shared with the externality of tools and artifacts, which are likewise employed by the brain on a general basis. As we want to distinguish the current discussion from the previous one about symbols and simulators, and given that external material objects cannot be integrated in these mechanisms in the same way  as language and symbols, we will presently focus on the question how we should interpret the hypothesis of 'extended cognition', particularly with regard to the interaction with these objects.

In their much discussed article on the 'extended mind', Clark and Chalmers do indeed scrutinize the interaction of humans with external objects like pens, notebooks, nautical slides, calculators. They propose to interpret this interaction as a case in which an external system becomes coupled in such a reliable and robust way to cognitive processes going on in the brain, that such objects significantly expand and alter these processes. Famous is the example of Otto, who must rely on an external memory in the form of a notebook that he always carries and uses. Otto is dependent upon this external memory in a way comparable to a normal person's reliance on his biological memory. If this is indeed accepted as a valid comparison, the authors claim that we may need to seriously reconsider our understanding of cognition and the boundaries of the system underlying cognition and behavior (Clark and Chalmers 1998).[255] While traditional accounts of cognitive processing tacitly maintain the skull as a fixed

---

[253] Clark refers critically to Dennet's idea of 'simulating' a 'more-or-less serial virtual machine' by the parallel hardware of the human brain with the help of language (Dennett 1993, 218 ff.). In a 1997 paper Dennett discusses some insights of Clark's (Clark 1997), yet maintains that language is a tool – yet a far-reaching tool – for thinking as it can influence the projects, rules, policies and so on of our thinking (Dennett 2000). Clark assigns an important role to language as an external prop complementary to the brain, thus leaving cognitive processes largely unaffected by language (Clark 1997 ; Clark 2008). In Clark's version, therefore, one could find a 'fixed properties' view of the brain (Kirchhoff 2012), quite different from our view of modifiable dynamical cognitive mechanisms. However, Clark increasingly appears to acknowledge the dynamic properties of the brain and their importance for extended cognition, emphasizing now that: "the biological brain adapts, selects, and alters, its own internal routines" for optimal exploiting external resources (Clark 2011 459). Indeed, in addition he contends that "the biological brain is the essential core element", suggesting a more moderate take on extended cognition.

[254] Clark refers to Barsalou's perceptual symbol systems (Barsalou 1999c), yet still stresses the issue of complementarity of such systems to the cognitive processes going on in the brain instead of recognizing how these symbol systems modify cognitive processing. He maintains that "language need not profoundly reorganize the shape and texture of the neural coding routines themselves" (Clark 2006 302), whereas Barsalou suggests that language – with its productivity, systematicity and capacity for recursivity – does have an impact on coding and re-activation of symbols (Barsalou 1999c).

[255] Indeed, according to that view, the mind is a plastic, open-ended system, "fully capable of including nonbiological props and aids as quite literally parts of [itself]" (Clark 2003 10).

boundary, it would then be more appropriate to refer to 'extended cognition' carried out by an 'extended mind' (Clark and Chalmers 1998).[256]

Our capability to integrate not only external information but also objects and technology in our cognitive and behavioral routines has been emphasized by Clark and others variously and in sometimes rather poetic statements.[257] Opposing what he considers outdated accounts of cognition that focus on our neural apparatus and tend to overlook or underestimate this capability, Clark contends that we humans: "make ourselves into new kinds of cognitive engine by (amongst other things) annexing and co-opting elements of external cognitive scaffolding as proper parts of hybrid computational routines" (Clark 2006 299).[258] Pens and notebooks were relatively simple, but we also see more recent examples of information processing and communication devices that we take for granted even though they vastly modify and expand the cognitive tasks we are able to perform, therefore deserving to be called 'pseudo-neural' (Clark 2003 45). Still more invasive are appliances like an additional,

---

[256] This argument is akin to Wilson's hypothesis of 'wide computation', which objects to methodological individualism in psychology, as there are at times environmental features that play a role in an individual's computational processes (Wilson 1994).

[257] Or in Clark's words, humans are potentially: "human-technology symbionts: thinking and reasoning systems whose minds and selves are spread across biological brain and nonbiological circuitry" (Clark 2003 3). Two objections have been made, both challenging the ontological status of the system that is assumed here. First, Clark's view is said to fail in appreciating that cognition does process content derived from social practices or conventions, like humans do when they process linguistic or mathematical symbols, and so forth. A second objection concerns the causal processes underlying cognition, that are different from external causal processes – some of which might be related to cognitive processes. Both objections are presented as grounds for rejecting the extension of mental or cognitive processes with causal processes of a different nature some of which involve derived contents (Adams and Aizawa 2001). Now the first objection is related to our seventh kludge characteristic, stating that a kludge can be established with the integration of environmental information. Derived content, can indeed become integrated in cognitive mechanisms, as is demonstrated when language switching in bilingual subjects implies alterations of their cognitive processes and the underlying neural networks (Garbin, Sanjuan et al. 2010 ; Kobayashi, Glover et al. 2008). However, the extended cognition hypothesis does not deny such diversity, as the hypothesis in fact emphasizes that hybridity regarding content and causality characterizes many of the cognitive processes that we usually accept as instances of cognition (Clark 2008). We would rather raise another critical remark. Once we recognize that cognitive processes are capable of kludge formation in which process environmental information and objects can play an important role, we still are not forced to include the latter into an account of the cognitive mechanism or system itself. Indeed, it seems to us that the extended view still needs to argue why ontological notions such as we find in Aristotle's Metaphysics (cf. the mereology given in V, 6), like coherence in time and space and the relative independent existence, are not applicable to the embodied cognitive system. To conflate two systems that are distinct in some basic ontological senses, even if tightly coupled, into a single system requires stronger arguments than those that merely show how each system by itself is already made up of different components. See for related mereological issues with regard to arguments in cognitive neuroscience our (Keestra and Cowley 2009 ; Keestra and Cowley 2011).

[258] In his critique of the extended view of the mind, Sterelny also refers to the process of external scaffolding, but considers this merely as another kind of 'niche construction', as we can witness many evolving species doing likewise by building nests or laying scent trails (Sterelny 2010). This account is more in agreement with ours, as it still focuses on cognitive mechanism's capabilities for kludge formation with the integration of environmental information but does not extend the mechanism itself out into the environment.

third – prosthetic - arm, a cochlear implant which allows ultrasonic sound processing, or an implanted retinal display directly connected to an external information database for processing the perceived environment, as these devices not only substitute our common capabilities but even add novel options for interaction with the environment. With these devices in place, we are allegedly no longer just conducting hybrid computational processes, but have become truly 'biotechnological hybrids' or 'soft-selves': "continuously open to change and driven to leak through the confines of skin and skull, annexing more and more *nonbiological elements as aspects of the machinery of mind* itself" (Clark 2003 137, italics added).[259]

Although the openness and modifiability of cognitive mechanisms has been at stake in this part all along, let us consider whether we can explain the interaction with objects, artifacts and technologies within the present framework, which emphasizes the formation of kludges with the integration of external information. Earlier, we noted that handling a tool may become integrated in one's performance, as when an expert singer finds that singing a difficult Don Giovanni or Saint François melody is facilitated by his manipulating a sword or cross. So yes, external objects can even shape cognitive processes, suggesting that these objects somehow play a role in complex functional mechanisms.

A relatively simple example of the modulatory influence of objects can be seen in monkeys, when the receptive field of visuo-somatosensory neurons expands after a limited period of handling an arm-extending rake, the tool apparently being integrated into its body schema (Iriki, Tanaka et al. 1996).[260] Further investigations supported this notion of the brain's capability to flexibly integrate external tools in the activation patterns of sensorimotor areas that were previously determined by the confines of the body (Mahon, Schwarzbach et al. 2010 ; Peeters, Simone et al. 2009).[261] However, this phenomenon is still of limited value regarding the more far-reaching conclusions

---

[259] Elsewhere, Clark identifies the body with the roles it plays in intelligent behavior, it being: "the locus of willed action, the point of sensorimotor confluence, the gateway to intelligent offloading, and the stable (though not permanently fixed) platform whose features and relations can be relied upon (without being represented) in the computations underlying some intelligent performances" (Clark 2008 207). The rejection of representation in this context is based upon a rather strict notion of that disputed concept. A more liberal interpretation of representation would allow for a fruitful employment of that notion in mechanistic explanations – see the discussion referred to in note 99 in chapter I.5 and elsewhere.

[260] The authors of (Iriki, Tanaka et al. 1996) interchangeably use the terms of body schema and body image when referring to this phenomenon of neural plasticity. However, to concur with Gallagher, it makes sense to use body schema when referring to: "a system of sensory-motor capacities that function without awareness or the necessity of perceptual monitoring" (Gallagher 2005 24), while awareness, attitudes and beliefs play a role in one's body image only.

[261] An issue that is yet to be clarified is whether birds or non-human mammals can create or discover new tools by employing some knowledge of physics, but to date there is little evidence for such sophistication in animals (Emery and Clayton 2009).

discussed here.

For the simple conclusion would be that tools like rakes and sticks merely co-opt the already existing neural representations of our body parts which are then flexibly adjusted – comparable to the reverse phenomenon when a neural representation of a body part in the brain is adjusted upon cutting a nerve to it (Buonomano and Merzenich 1998). The arm extending tools appear to modify the explanatory mechanisms involved in reaching, touching or perceiving only superficially, by adjusting relevant sensori-motor parameters.[262] Several anatomical studies have shown how as a consequence of tool use the areas of both somato-sensori neurons and visual neurons are enlarged, in this way extending the peri-personal space of the monkeys. PET scan research in humans, using not only rakes for reaching but also tongs for picking up objects, shows comparably enlarged activation sites (Maravita and Iriki 2004).[263]

However, tool use with a larger impact than just dimensional adjustments has been studied as well, potentially offering a more convincing demonstration of our being 'human-technology symbionts' in Clark's sense (Clark 2003).[264] With more complex tools, novel actions are made possible involving action means or goals that are impossible to obtain with mere bodily movements or their elongated versions. Experiments with monkeys and humans who learned to use both normal and reverse pliers (that allow grasping an object when opening instead of closing the handles)

---

[262] It is possible to disentangle the sensorimotor adaptations that occur with tool-use and the modification of the feedback that is provided by the 'distalized' visual target (Arbib, Bonaiuto et al. 2009). In similar research with rubber hands, subjects report illusory sensations of being touched when they perceive how a prosthetic hand or rubber hand is being touched, if these are carefully positioned and manipulated by an experimenter relative to their body or attached to it (White, Davies et al. 2010). This fact demonstrates that these objects are not merely used as external instruments but indeed are integrated in a multimodal body schema. Interestingly, patients suffering from schizophrenia are more susceptible to such illusions, interpreted as failures in processing body ownership. Such failures are relatively easy to explain, as they rely on failed detection of the incongruence between visual, tactile and proprioceptive signals (Jeannerod and Pacherie 2004 ; Thakkar, Nichols et al. 2011).

[263] Holmes and others argue that there may at least be an additional role for attention allocation in the explanation of the phenomena reported about tool use and the extension of an agent's peri-personal space, as that phenomenon seems to be vulnerable to interferences with attention (Holmes, Calvert et al. 2004). As such acquired neural extensions are not as stable as the projections that stem from body parts, one could indeed expect these neural extensions to be more vulnerable, indeed. Nonetheless, investigation of the changes in visual-tactile representations along the dimension that became extended with tool use shows that these changes occur in a gradient along that axis, ruling out that it is only a matter of attention allocation (Bonifazi, Farne et al. 2007).

[264] An indication of the importance of the function of a tool rather than its mere dimensions in 'sculpting' the agent's peri-personal space is provided by an experiment where a 60-cm. stick is used, with a handle positioned halfway at 30 centimeters. The extra, yet useless, 30 centimers did not extend the peri-personal space, confirming other research that demonstrated the importance of the agents' active involvement with a tool's functions for neural and cognitive adaptation to the tool (Farne, Iriki et al. 2005). Indeed, only when agents intend to use the tool for reaching an object, their peri-personal space was found to become actually extended. The authors point out that both ability and intention to perform an act modulate perception in these cases (Witt, Proffitt et al. 2005).

show that movements and action goals are differentially coded in the brain, allowing flexible configurations (Cattaneo, Caruana et al. 2009 ; Umiltà, Escola et al. 2008).[265]In that sense, such novel actions would be better comparable to the novel situation simulations that are made possible by the configuration of endless, new and abstracted combinations within perceptual symbol systems (Barsalou 1999c), as we discovered in the sections above.[266] Indeed, the interaction with tools or artifacts has been shown to: "change the way the human brain perceives the size and configuration of our body parts" (Malafouris 2010).

Now it is well known that premotor and parietal areas are activated not only by motor engagement with objects and tools, but also by the observation, the imagination or planning of potential motor actions (Jeannerod and Frak 1999). Apparently, motor representations do not just represent the complete, stored experiences of complex motor actions. On the contrary, these motor representations are stored in such a way that they allow the explicit or tacit recomposition of such actions in order to facilitate the prediction of potential outcomes when interacting with the environment (Jeannerod 2006).

Adding another layer of complexity, the brain is usually composing several potential action representations in parallel, requiring a selection of the single action that is eventually performed. All in all a complex task, involving a highly distributed network of parietal, motor and also prefrontal cortex areas (Cisek 2006). The integration of external objects and tools in potential actions does then require some extra – but not completely novel - adaptivity and flexibility, added to an actor's expertise of tacitly composing action representations for bodily actions with familiar means and goals. Humans appear to be particularly good at this.

Indeed, studies of (over-)imitation in humans and animals that use 'artificial fruit', or opaque boxes that have to be opened via complex sequences of actions, demonstrate that humans are inclined and better able to observe, analyze and compose comprehensive representations in order to imitate complex actions and object use.[267]

---

[265] Other research shows how handling a contemporary tool like a (manipulated) computer mouse similarly leads to novel representations in the brain. The authors demonstrated the presence in the cerebellum of multiple, modularized models of features of novel mouse actions, rather than models of the hand actions that were required for using the mouse (Imamizu, Kuroda et al. 2003).

[266] The limitations in nonhuman primates in tool use may be related to their demonstrated inability to master hierarchical structure in language (Fitch and Hauser 2004), as both require complex computations and coordination at several levels of complexity. Study of gestures in gorillas show an equally limited repertoire with very little evidence of the invention of new gestures for novel situations or forms of idiosyncrasy (Genty, Breuer et al. 2009). Cognitive archaeology similarly suggests that tool use – primarily in the form of stone use and stone knapping – emerges well before the development of symbolic representations or of language (Malafouris 2010 ; Stout and Chaminade 2009).

They handle such representations at several hierarchical levels of specificity whereas animals tend to leave out intermediate steps and focus on overarching action goals only (Lyons, Young et al. 2007 ; Nielsen and Tomaselli 2010). Such differences have caused wonder, given the fact that there are clear homologues between monkeys and humans regarding their neural systems involved in reaching, grasping and manipulation of objects. This has led to speculation that perhaps monkey limitations in causal knowledge are responsible (Johnson-Frey 2003). Apart from that possibility, there is abundant evidence for the elaborate neural representation of actions in humans, which allows humans the elaboration of a much more extensive action hierarchy than animals (Badre 2008 ; Grafton and de C. Hamilton 2007). Such a representation of an action at several hierarchical levels enables a flexibility that also strongly supports the integration of familiar or novel tool properties– potentially including their causal properties and relations.

Associated with this flexibility is the cognitive task of discovering whether or not an object allows for a particular action, which requires the perception and representation both of an action goal associable with the object and of the behavioral steps required to reach that goal (Masson, Bub et al. 2011). More complex still are situations in which tool use involves adjusting a familiar object to a novel application, as was the case with the use of reversed pliers mentioned earlier, which was associated with modified action representations and underlying neural activations (Cattaneo, Caruana et al. 2009).[268] In order to become the veritable 'human-technology symbionts' (Clark 2003) humans are said to be, it is important that humans gain the same sort of expertise in adapting and recomposing action representations with objects as they have for actions without objects.

Now it is predominantly the prefrontal cortex that supports the establishing and learning of associations between various perceptual and motor representations necessary for the configuration of complex actions involving external objects (Fuster 2000). Interestingly, comparative evidence suggests that the prefrontal cortex partly

---

[267] This finding appears in contrast with evidence taken to demonstrate 'rational imitation' and teleological reasoning capacities in young children. They appear not to imitate an irregular use of the head for switching on a light if they can infer from the situation that it is equally permissible to use their hands (Gergely, Bekkering et al. 2002). However, in a review of overimitation research in which rational imitation is discussed as well, the authors point out that overimitation occurs primarily in cases of tool or artifact use, where the opacity of the means-ends relationships precludes such immediate conclusions about the (ir)relevance of the actions involved. Overimitation thus helps children to find their way in our 'artifact-centric culture' (Lyons, Damrosch et al. 2011).

[268] Neuroimaging investigations of skilled performances with different tools - as with two computer mice with contrasting properties - show that the two skills activate different cerebellum locations, suggesting that different representations are formed for each skill (Johnson-Frey 2004). Given the large representations of skilled actions, there will also be overlapping components of these kindred representations.

evolved as an extension of the motor cortex, which would explain why it is so much involved in action control and why it enables increased capability of learning to compose and control complex actions, including tool use. Differences in tool use between chimpanzees and bonobos correlate, for example, with the size of their dorsolateral frontal areas (Stout 2010). Expertise in tool use is shown to have similar learning effects as expertise in common motor actions in an imaging experiment with humans who were observing tools being used in common and uncommon manners. Distinct effects were observable for expertise with goals and with the means of actions (Valyear, Gallivan et al. 2012). Such evaluation and selection between multiple action representations, implied in the adjustment of action representations, is supported by large and distributed representations of action features, which require prefrontal cortical activations in humans (Cisek and Kalaska 2010). Just like Barsalou's simulators are associated with complex and distributed representations, it appears that a similar type of representation underlies action and allows its composability and versatility.

Given such confirmations of the brain's capability to develop new action representations with the integration of new information about external objects, it does not come as a surprise that the ability to learn to control high-tech appliances can even be found in monkeys. Indeed, monkeys demonstrated fast learning to control a virtual grasping hand through brain-machine interaction. The brain-machine interaction involved both movement control with electrodes connected to the monkey's primary motor cortex and some tactile feedback by electrodes connected to its sensori areas, not very different from normal tool use (O'Doherty, Lebedev et al. 2011). Such insights suggest that 'human-technology symbionts' (Clark 2003) probably differ more in degree than in kind from their 'monkey-technology' counterparts. Moreover, they confirm our notion that brains are generally capable of developing complex routines, taking into account relevant external information and properties of objects and artifacts.

So far, this part has focused on the dynamic nature of the complex mechanisms that underlie our cognition and behavior, highlighting the phenomenon of kludge formation which is partly responsible for shaping and reshaping the space of actions available to any agent – whether or not in symbiosis with technology. Let us now check whether kludge formation can account for some of the astonishing phenomena that the extended cognition hypothesis asks attention for.

## 4.4  Cognition-extensions and the kludge characteristics

The article that started the discussion of extended cognition emphasizes the functional properties of many cognitive extensions while defending the non-neural nature of these extensions, concurring with our first kludge characteristic which refers to it

being predominantly recognizable in functional properties. Consequently, it was defended to recognize any process as cognitive because of its functional properties, even if it is carried out by a part of the world and not inside the head (Clark and Chalmers 1998). This so-called Parity Principle is testimony of the fact that the extended cognition hypothesis stresses the functional continuity of processes going on in the embodied brain and in its environment. Indeed, Clark asks us to: "judge various potential cognitive extensions behind a kind of 'veil of metabolic ignorance'" (Clark 2011 449), implying that differences in physical implementation of the cognitive (component) processes should not bear much weight. As is to be expected, this strict focus on the functional properties has been challenged as neglecting an adequate role for the physical system and its boundary conditions that would be responsible for these functional properties: these conditions are usually different for the brain and for some of its extensions or the technologies it uses (Rupert 2010). Indeed, as soon as we consider the involvement of external objects, artifacts, or technologies in the functional properties of extended cognition, problems arise. For example, it is difficult to see how the process of Representational Redescription that is often involved in development and learning (Karmiloff-Smith 1992) can have an impact not just on the cognitive processes involved in tool use, but also on the tools or cognitive extensions like notebooks and calculators themselves. It appears that Clark also recognizes that irrespective of its potential cognitive extensions, the main function of cognition remains the same, when he observes that the: "overarching goal of minimizing informational surprise can be served (...) by the canny longer-term structuring of an environment" (Clark 2011 454).[269] Thanks to the brain's capability of kludge formation while integrating in multiple ways the relevant properties of external objects and artifacts, we are capable of cognitively processing information that would otherwise have remained impossible. Even though it is now agreed that it is: "the biological brain [that] adapts, selects, and alters, its own internal routines so as more and more fluently to exploit the reliable presence of all those specific, culturally selected, tuned, and delivered, resources" (Clark 2011 459), these resources allegedly modified and expanded the brain's cognitive capabilities.[270]

The second kludge characteristic focuses on the algorithmic theory that could

---

[269] This observation is associated with the thesis that the brain is a 'prediction machine', continuously involved in anticipating future perceptions or actions based upon previous experiences (Clark 2013). Similarly, it has been argued that the brain aims to minimize prediction errors (Friston 2010), or is proactively anticipating the future on the basis of past and current experiences (Bar 2009). Common to these theses is the flexible involvement of complex, distributed representations, too.

account for the kludge formation – in the present context potentially including representations pertaining to cognitive extensions. Since it is an essential feature of extended cognition to be hybrid both in terms of the recruited resources and in terms of the information to be represented, the question is whether there is a particular algorithmic theory involved and if so, whether we can determine it. In the first account of extended cognition, it was language in particular that was considered to be the tool that has: "the major burden of the coupling between agents", allowing us to "spread this burden into the world" (Clark and Chalmers 1998 18). If language plays such a central role, we might use linguistic knowledge to derive some very general features of an algorithmic theory that is associated with a particular kludge. But what happens when the hybridity of extended cognition extends to other options for the representation and computation of information, like when representations are used that are geared to our sensorimotor capacities, like levers, movements, and so on (Clark 2008)?[271] Intriguingly, it might be easier to develop algorithmic theories for such cognitive extensions than for language-dependent ones. For example, where it is principally impossible to reconstruct with certainty most cognitive strategies of persons living in antiquity, it may be easier to explain their cognitive extensions: their symbol systems and their arithmetic devices. In all cases we might expect that patterns of stability comparable to the stability that ensues upon kludge formation have been developed, but it is often easier that reconstruct the representations – implied in Marr's algorithmic theories - describing the interactions with these objects are usually much constrained by the highly determined demands and affordances of these. For this reason, extended cognition often amounts to the emergence of 'horizontally extended cognitive modularity' (Wheeler and Clark 2008).[272] For example, the functionality of normal and of reverse pliers is easy to recognize from

[270] An important question remains how we apply the Parity Principle. For example, it is difficult to see how Otto's notebook would be continuously and effortlessly updated once new knowledge about the museum's location or collection is obtained during an accidental discussion (cf. Clark and Chalmers 1998). In the biological brain, on the other hand, there is ever more evidence that the so-called 'default mode network', an identifiable network that becomes active in the absence of actual cognitive or motor demands, is precisely responsible for such maintaining and updating of information (Raichle, MacLeod et al. 2001 ; Raichle and Snyder 2007). Indeed, given that the integration of information is an important and complex cognitive task, such a network would play indeed a crucial role (Hohwy 2007). Updating external resources would require extra efforts, time and attention, which implies a transgression of the Parity Principle, it seems.
[271] Clark builds also on Wilson's notion of 'wide computation', which involves external resources and alternative information structures like pen and paper and mathematical notations. Wilson focuses particularly on an explanation of computation and argues in that context for a non-individualistic account (Wilson 1994). Compared to that computational focus, Clark's ambitions are much larger as they pertain to an analysis not just of computation, but of the human mind in general, as testified by his book titles "Natural-born cyborgs. Minds, technologies, and the future of human intelligence" (Clark 2003) or "Supersizing the Mind. Embodiment, action, and cognitive extension" (Clark 2008).

their appearance, each provoking specific neural activation patterns. Encoding of the specific movements afforded by the handles occurs separately from the encodings of the action goals enabled by the pliers (Cattaneo, Caruana et al. 2009 ; Umiltà, Escola et al. 2008). Obviously, not all tools are equally transparent and opaque artifacts or tools will not allow observers or users to infer their functionality and handling (Lyons, Damrosch et al. 2011). Still, it is probable that at some level of specificity we may determine the algorithmic theory of extended cognition even better than for cognitive processes without such extensions.

At times, we may be able to derive an algorithmic theory of a kludge somewhat easier in cases of tool use than in previously discussed cases of kludge formation. Does this also hold for our third characteristic, concerning the neural implementation theory? Originally, the hypothesis of operations performed by the extended mind where considered to be the result of two distinct yet coupled systems, to wit: the cognitive system and an environmental system or object (Clark and Chalmers 1998). However, over time the argument has emphasized consideration of a single yet complex system as the source of such operations, with a "complex cognitive economy spanning brain, body, and world" (Clark 2008 217). Obviously, that system will involve not just a neural implementation but also corresponding bodily and environmental implementations. An important difference with the kludges discussed in earlier chapters is the fact that the kludges involved in this economy tend to be less stable because of these complexes being 'soft assembled', making them "transient extended cognitive systems" (Clark 2008 158). Moreover, their presence is dependent on multiple and different kinds of conditions, like those on which the add-ons like instruments and pen and paper depend. Compared to the systemic integrity of the brain – which is not soft-assembled from situation to situation, integrating environmental objects that are available – such extended cognitive systems can be characterized by the presence of more sets of highly different constraints and limitations, related not only to the embodied brain but also to those objects and the interactions with these. As a result, such systems are generally much more vulnerable. Indeed, this lack of systemic integrity is for Rupert a reason to distinguish principally between extended cognition and non-extended cognition (Rupert 2009).[273] Similar but with a different emphasis, it is argued that the complex nature of these extended cognitive systems is characterized by more

---

[272] In their account Wheeler & Clark build on the neuroconstructivist accounts of Karmiloff-Smith (Karmiloff-Smith 1992) and others (Mareschal, Johnson et al. 2007), because these demonstrate that the emergence of functional modularity can occur during individual development and learning and is not just a result of evolutionary processes (Wheeler and Clark 2008).

than just a single translational input-output connection within the system. That is, whereas normal cognitive processes involve representations derived from physical sensory input, leading to physical motor output, in extended cognitive systems there are many more boundaries present where such transduction occurs. This is the case, for example, when cognitive processing relies partly on additional information representations, as when these are written down in a notebook, as in Otto's case (Weiskopf 2010). Indeed, as much as tight and stabilized interactions between cognitive processing and external information or objects is possible, considering the latter to be equally constitutive components of a cognitive system as components of the embodied brain seems an overstatement (Aizawa 2010).[274] Even though extended cognition is comparable to some extent with kludge formation and particularly relies on the capacity for kludge formation, the implementation of an extended cognitive system differs significantly from the implementation of a kludge in the brain.

Differentiation between general kludge formation and the formation of kludges in interaction with cognitive extensions brings us to the fourth kludge characteristic, pertaining to the variation between stages of kludge formation. In our earlier discussions of variability during kludge formation, its stability and modifications over time were an important issue. In previous sections of thie chapter we already dealt with kludge formation due to increased expertise with language and symbols, so we will here focus on the variations in the use of tools. In cases where environmental objects are used as tools – like the notebook and pen upon which Otto is dependent for much of his actions (Clark and Chalmers 1998) – this variability is not only dependent upon (embodied) brain processes, but is also dependent upon the wear and tear of those objects, on the weather conditions that may affect them, their replaceability, and so on. Independent of the level of expertise with particular tools,

---

[273] The emphasis on the integrity of the cognitive system does not imply an underestimation of the role of external instruments or information for cognitive processing. On the contrary, Rupert argues: "the fundamental theoretical construct of virtually all successful cognitive science – whether computationalist, connectionist, or dynamicist— is that of a persisting architecture interacting with an ever-changing cast of external materials to produce intelligent behavior" (Rupert 2010 344). Such an observation regarding the involvement of external information or objects in cognitive processing is not new, of course. A comparison of the hypothesis of the extended mind with Hegel's idea of an objective spirit that supports individual cognitive processing demonstrates also some overlap in those analyses, for example (Crisafi and Gallagher 2010).

[274] The determination of the boundaries of explanatory mechanisms depends upon the consideration of the 'constitutive relevance' of the component parts and processes involved. Craver appears to concur with the hypothesis of extended cognition that at times cognitive mechanisms 'draw upon resources outside of the brain and outside of the body to such an extent' that we may recognize these to be constitutive of these mechanisms (Craver 2007 141). Although Craver draws upon the notion of robustness as developed by Wimsatt (Wimsatt 2007), he is less careful in distinguishing between levels of robustness.

an extra variability is involved as soon as components with such different properties are assembled to form a kludge. The differential variability of the components of the extended cognitive system is particularly visible at the interfaces that we generally can localize in such a system, for example when an artist has great control over his hand, less so over the pencil in his hand, but probably less over the interaction between pen and paper (Rupert 2009 cf. p. 170). However, once we restrict our focus to those cognitive processes that are involved in tool use, we can observe again different levels of learning and expertise with corresponding changes in neural activation patterns (Valyear, Gallivan et al. 2012), suggesting that kludge formation is involved here, too. And indeed, mastery of a tool consists partly in the expertise with handling the tool's oddities and compensating for these. As a result of that, an expert tool user will demonstrate more stability in his performance with the tool than a novice will (Charness and Tuffiash 2008), again confirming the impact of the process of kludge formation.[275] This does not erase, however, the fact that with extended cognition, agents have to cope with multiple sets of constraints that are valid for the various components of extended cognitive systems. As a result, we may observe more complex and differential patterns of variability, depending on the stability of the distinct components with their quite different properties.

From this diversity and variability of the components of kludges that emerge from extended cognition, there is a straight connection to the fifth kludge characteristic. When we mentioned earlier a kludge's 'cobbled together' nature (Clark 1987 291) it referred mainly to the neural and cognitive processes out of which a kludge is developed. In the case of extended cognition, we are looking at a wider range of resources. Obviously, this earlier notion is still valid in this context. Indeed, in Clark's early article on 'the kludge in the machine', he emphasizes that complex human cognitive processes are not completely novel but are grounded in and built from the "proto-cognitive capacities we share with lower animals" (Clark 1987 291). Gradual changes to a cognitive system with such capacities can have large snowball effects given enough time. Once one agrees to include into the system environmental resources as intrinsic components, the composite nature discussed here is even more obvious. Indeed, Wheeler & Clark argue that the human cognitive system is intrinsically open and is in fact: "a cognitive machine intrinsically geared to self-transformation, artefact-based expansion and a snowballing/bootstrapping process of computational and representational growth" (Wheeler and Clark 2008 3572).

---

[275] Indeed, differences in activation patterns between experts and novices are also found when subjects are merely imaging the use tennis rackets, showing extended neural activations only in experts (Fourkas, Bonavolontà et al. 2008).

Above was mention of the capacity of the brain – of the expanded human brain in particular – to compose complex and flexible representations for interacting with the environment with the integration of external information in these representations (Cisek and Kalaska 2010). Instead of considering such interactions as demonstrations of: "the nonneural body and the nonbodily world as each capable of making key cognitive contributions" (Clark 2007 164), it is defended that this is just another example of niche construction, it being continuous with many other such niche construction phenomena in nature (Sterelny 2010). Such defense implies that calling these external resources 'not alien but complementary' to the brain's contributions to extended cognition (Clark 1997 ; Clark 2008) is still underestimating the importance of the fact that these external resources are not produced by neurophysiological processes prevalent in the body and brain, making it relatively easy to distinguish them from neural components. Given the relative independence of the neural components of cognition from their cognitive extensions, it is not plausible to put the former on a par with the latter regarding their involvement in cognitive developments (Adams and Aizawa 2001 ; Rupert 2009).[276] This is in contrast with the prevalent re-use which happen to neural structures (Anderson 2010), as was the case with modularized neural networks involved in child learning (Karmiloff-Smith 1992), with networks involved in automatized processes (Frankish 2010) and with simulators when they have emerged (Barsalou 1999c).

Discussing the sixth kludge characteristic seems less problematic than the previous ones, as it is part of the hypothesis of extended cognition that cases of 'horizontally extended cognitive modularity' (Wheeler and Clark 2008) or of hybrid thoughts or hybrid computational routines (Clark 2006) can become involved in further cognitive developments or trajectories. From his earlier analysis of 'gradual holism and the historical snowball' (Clark 1987) onwards, Clark has emphasized how earlier cognitive developments both open up and constrain the option space for later ones. What remains important in the present context of extended cognition, however, is

---

[276] Responding to Rupert's worries, Clark insists on the Parity Principle which focuses merely on the functional comparability of the bodily and environmental contributions to cognitive processes with those contributions that are going on in the head. Moreover, Clark concedes that it may be a difference in grain of their respective analyses that is partly responsible for their dispute, as the Parity Principle does not require 'fine-grained identity of causal contribution' (Clark 2007 168). In terms of mechanistic explanation, the analysis of this dispute looks different: after a first decomposition of a cognitive phenomenon, scientists will look for the mechanistic components that are responsible for the phenomenon or its phenomenal components. It may well be that in doing so, the lack of robust interactions or coherence between some component parts and operations or the presence of several distinct loci of control is such that it makes more sense to explain the (cognitive) phenomenon as the product of several – interacting, perhaps coupled – separate explanatory mechanisms. Mereological considerations matter here, again (see note 257).

the fact that not all components involved share the same scales of time and space for their development or are equally capable of development. Physical components like the eye or the swim bladder develop gradually during the course of evolution, affecting eventually whole species and their further evolutionary trajectories (Clark 1987). In contrast with those, it is more difficult to predict the 'epidemiology' of the cognitive representations or material artifacts involved in cultural developments (Sperber 1985). For example, cultural components of extended cognition may transform within only a single generation, but depending upon their representational format – in a specific language, for example - these form the environment of a community that is rather limited in its geographical distribution. As much as the authors describe extended cognition as a consequence of human nature's 'extensive openness to training and input-based modification' (Wheeler and Clark 2008), the authors unfortunately pay scarce attention to the differences in generative entrenchment between natural and cultural components – with objects presenting yet another class - and between their developmental trajectories. For when kludges are established, for example in cognitive mechanisms, and are then involved in further developments, as a consequence the previously established kludges are becoming deeper generatively entrenched in the organism. That is to say, it may be possible to differentiate between older or more foundational entrenched kludges and those of a more recent and superficial nature, building upon those older ones (Wimsatt 1986). Consequently, if the integrity of such a foundational kludge is being compromised this is likely to have a chain of effects, as when someone's language skill is compromised and his overall social and even cognitive functioning is disturbed. Applying this notion of generative entrenchment in order to differentiate between the various trajectories of kludges in biology and culture, might offer a highly welcome nuance to the hypothesis of extended cognition [cf. (Wimsatt 1999 ; Wimsatt 2001 ; Wimsatt 2006b ; Wimsatt and Griesemer 2007). It would again confirm Rupert's emphasis upon the lack of integrity of the system that underlies extended cognition by (Rupert 2009), given that there are great differences between the components of extended cognition in this respect. Even though comparative evidence, including archaeological evidence, suggests that tool making and the development of language capability have co-evolved in humans, human tools tend to change more rapidly and contingently than language does – while the brain's evolution occurs at an even lower pace (Stout and Chaminade 2009). So even when the two differ in their influence on brain evolution, language and tool use are comparable with regard to the complex and composed representations associated with these and which are differentially involved in subsequent developmental and cultural trajectories (Roepstorff 2008).[277] Unfortunately, in his discussion of 'material symbols' Clark only

refers to linguistic and mathematical symbol use and shuns the question whether the material properties of tools can become equally integrated in human cognition and action only on the basis of their *cognitive* representation, which is what we are arguing here (Clark 2006). Nonetheless, our view appears to concur with Clark's more recent emphasis on the modification of the brain's 'internal routines' as the condition for exploiting external resources (Clark 2011). When an individual develops such routines in tool use and these give rise to kludge formation, they will alter the option space of his further developmental and learning options.

With that we arrive at the seventh kludge characteristic, referring to the involvement of external, environmental information in kludge formation. That external information is important for the hypothesis of extended cognition is no longer a surprise. Nonetheless, the term 'external' is somewhat difficult to define with respect to extended cognition, because cognition itself 'leaks' into the body and the world according to this hypothesis (Clark 2008). Setting aside the at times somewhat hyperbolic rhetoric, we've already noticed that the hypothesis does in the end not reject the perspective of cognitive science in putting the brain central to its considerations. Indeed, it is recently argued that this hypothesis concurs with a 'neurally-unifying predictive coding framework' according to which it is especially the brain's efforts to minimize informational surprise that unify all processes of extended cognition (Clark 2011).[278] Still, environmental information and objects can and do play a crucial role in this task because of human nature's inherently 'extensive openness to training and input-based modification' (Wheeler and Clark 2008), as we learnt above. Constructing their own environments, including the objects that occupy these, humans are in fact constructing niches that maximally employ this openness to their modification at several levels of specificity (Sterelny 2010). The formation of kludges in the interaction with language and tools and the involvement of these kludges in subsequent developments leads to an ever greater entrenchment of external information in the individual's cognitive mechanisms, in socio-cultural structures and perhaps eventually in the evolving brain of humans generally.

---

[277] In his review of comparative evidence concerning the development of ever more complex tools by humans, Ambrose does hypothesize that particularly the expansion of the frontopolar part of the frontal lobe in humans is driven by tool use. Basically, though, he emphasizes how language and tool require similar cognitive capacities for handling hierarchically structured representations: "Assembling techno-units in different configurations produces functionally different tools. This is formally analogous to grammatical language, because hierarchical assemblies of sounds produce meaningful phrases and sentences, and changing word order changes meaning" (Ambrose 2001 1750).

[278] Several authors argue that the brain's main task is to engage in such predictive coding of anticipated inputs in its engagement with the environment (Friston 2005), as this task of the – 'proactive' - brain enables optimal interactions of the organism with that environment (Bar 2009).

## 5  DYNAMIC MENTAL MECHANISMS, KLUDGE FORMATION AND ESTABLISHING CONSTRAINTS ON THE SPACE OF OPTIONS

This part started with the example of a whining and babbling baby that increasingly gains control over its voice and subsequently over linguistic and musical structures that eventually enable it to sing. We analyzed this trajectory of development and learning as a process in which kludges are formed that modify the mechanisms responsible for the child's functions. Building on insights from chapter I.5, this analysis should convince the reader of the fact that mechanisms which are used to explain cognition and behavior are highly dynamic and modifiable. Involving structural modifications as a result of development and learning, these mechanisms can obtain novel properties and capabilities. The phenomenon of a child who quickly learns to control complex functions and in a short time acquires many completely novel capabilities already supports this modifiability. It was confirmed once more at the end of chapter II.4, when we met the hybrid phenomenon of extended cognition, with biological brains effortlessly co-opting contemporary technologies. The latter phenomenon, however, emphasized another issue that we already observed along the way.

Mastering a skill or gaining expertise in any cognitive or behavioral function corresponds with a process of kludge formation, affecting the action space of these functions, so we argued. Some linguistic structures will be better mastered than others, some melodies sung automatically while others still require attention of the singer. Moreover, we also observed that such an action space pertaining to a function is not well-defined but can be expanded, covering areas that were previously separate. Indeed, the action spaces of previously distinct functions can become strongly associated or to some extent integrated thanks to the processes discussed here. Modularization of a particular neural network, for example, corresponding to development and learning of a specific skill, facilitates their activation, restricts the recruitment of necessary neural resources and diminishes influences from other neural networks. As a result, the demands on the brain decrease when the skill needs to be performed. Importantly, this not only leaves extra room for simultaneous performance of other cognitive or behavioral functions, it also facilitates the further development, elaboration and expansion of that particular skill. The child, for instance, may not just master the distinct skills of voice control, rhythm control, syntax and semantics, but could eventually integrate these skills enough to become an opera singer. With many routines being developed, some covering a highly general

and others a more specific domain, this child's responsible mechanisms have become ever more complex. This complexity is due to its simultaneously establishing ever more kludges and to the fact that some of these become strongly associated, perhaps even merging into a single, complex kludge.

As is generally the case in cognitive neuroscientific explanations, developments like these can be described at different levels of analysis or from different theoretical perspectives. In Part I we argued that both Marr's three levels of analysis as well as levels of mechanism would need to be taken into account for a comprehensive explanation of cognition and behavior. In both cases the notion of representation is useful, for example when an explanation of information processing is at stake or when we aim to explain how a particular motor action is modified, expanded, or the like. For our purposes, we especially made use of the notion of representation to clarify that kludge formation is not just a modification of processes going on in an isolated brain. On the contrary, so we argued, learning and development occur in continuous interaction with bodily processes, that also affect in specific ways the modifying brain processes, including the kludges that are established. Moreover, information pertaining to properties of the environment plays a role in these modifications and kludges. In some sense, then, environmental information becomes represented in the relevant action spaces – facilitating the interaction with specific environments, enabling greater speed and flexibility in responding to particular environmental stimuli.

A further observation concerned the fact that this entrenchment of environmental information in modifiable mechanisms and in kludges can have far-reaching consequences. Given the fact that a snowball effect occurs when a particular cognitive or behavioral process becomes automatized due to kludge formation, the entrenched information will affect subsequently activated processes as well. Cultural peculiarities in tone formation or pitch in speech, for example, will continue to influence a novice's singing. Only by paying due attention to this and with careful training – that is, by recruiting extra resources – may an expert regain control of such basic vocal functions and add different kinds of tone formation to his vocal palette. In so doing, he aims to establish more than a single kludge so that he can sing in different vocal modes. Indeed, when an expert singer has established kludges for several components of the mechanism involved for vocal control, it becomes easier for him to focus on the more subtle differences that characterize German or Italian. Although his singing is largely automatized, this does not preclude the expert's capability of having access to those subtleties that he normally would not pay attention to.  The representation of environmental information in such capabilities

and their development can apparently have very different forms and allow different modes of access. Still, even the representation of highly complex information, like mathematical structures or difficult atonal opera lines, can be involved in kludge formation and affect several interrelated cognitive and behavioral functions.

A final point that merits attention in this concluding chapter of Part II, as it will reappear in the third and last part, concerns the complex and composite nature of the representations involved in most cognitive and behavioral processes. The composability of words, sentences and stories is not something unique to human language, but has analogies in action and cognition, too. Indeed, this similarity in composability or configurability of the representations and activities involved in cognition, language and motor action is supported by mechanisms in the brain and rely on the modifiability of mechanisms which are not completely specific to humans. Irrespective of this fact, we found that particularly in humans there is an immense space of options available for such configuration of actions, for example. Since actions are not represented as single units but rather as composed of different components represented in a distributed way in neural networks, agents can form novel actions even without learning new components by reconfiguring familiar actions. The human brain, characterized among other things by its large prefrontal cortex, is particularly well-equipped for the representations necessary for complex action and cognition. Facilitated by kludge formation that may include particular components of well-learnt actions, an opera singer may in addition to his singing easily learn to accompany his singing with fencing with his sword or dancing with his beloved – a combination of actions that is far beyond the reach of a novice. Indeed, the complex and distributed representations involved in such complex actions offer many potential points of contact with external objects or persons that have to play a role in these actions. The configuration of an action can include an object as a component for which to prepare a specific manual grasping movement, as an instrument for hitting a particular structure, or even as the final goal of a particular sequence of actions. Depending on where such an object figures in the complex action, an agent needs to take into account its particular properties and integrate them in his action representation. Fortunately, again, such interaction can also become facilitated by kludge formation with the integration of relevant environmental information.

Having articulated these main lessons from Part II, we can now take the next and final step. Inspired by Aristotle's comparison between moral action and musical activities and with our insights from Part I, we will reflect upon complex intentional action – including actions that are subject to moral requirements. This comparison invites questions concerning the requirements for complex action, particularly

complex moral action. Do we not always expect moral action to depend upon complex and conscious decisions or intentions, making them very different from the automatized actions of a musician or singer? Indeed, is automatization not at odds with the adaptivity and flexibility that we usually associate with morally adequate action? Or does automatization rule out the accountability and responsibility for automatized actions, even though these are crucial features of human interactions? In sum, is the concept of a sculpted action space due to kludge formation at all applicable to intentional and moral actions, as we tend to associate such actions with unique decisions that take into account multiple kinds of information and can therefore not even partly be automatized? As may be expected from the foregoing, our argument will belie these assumptions. Don't we expect  from an expert singer both a nuanced and moving interpretation of a dramatic area and the capability to effortlessly respond to the conductor's and director's desires and the whims of his partners on the stage? Aren't we capable of recognizing his personal style in his performances, while still admiring his flexibility in adapting to the different contexts an expert has to deal with? Apparently, the singer is able to flexibly configure his complex behavior and differentially integrate environmental influences in it while still keeping to some of his personal long-term commitments and intentions. It is this fascinating phenomenon that will occupy us from now on.

# Part III
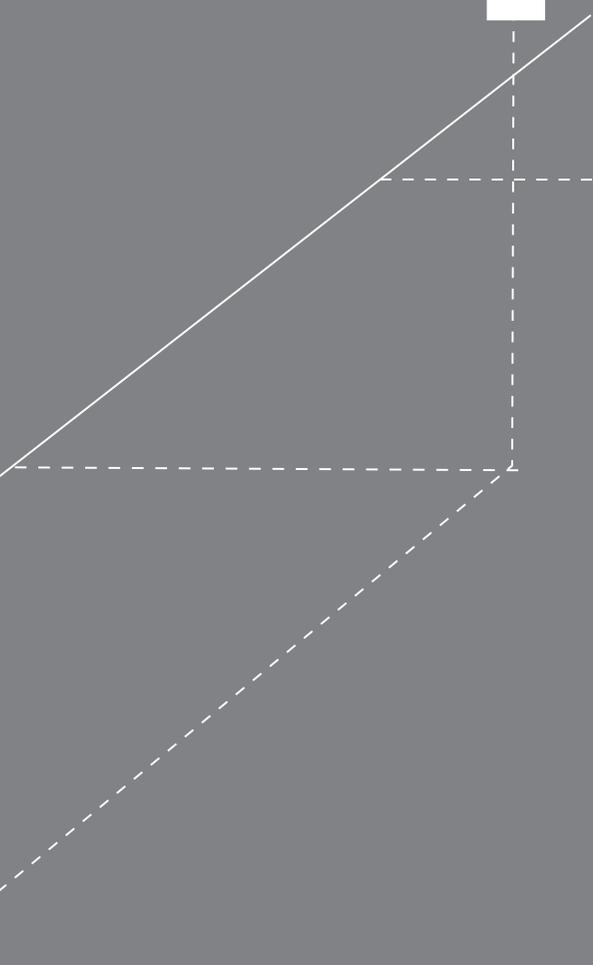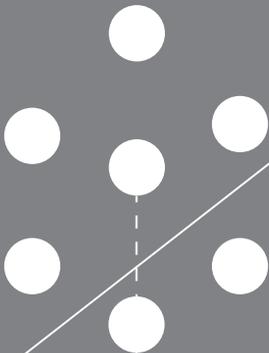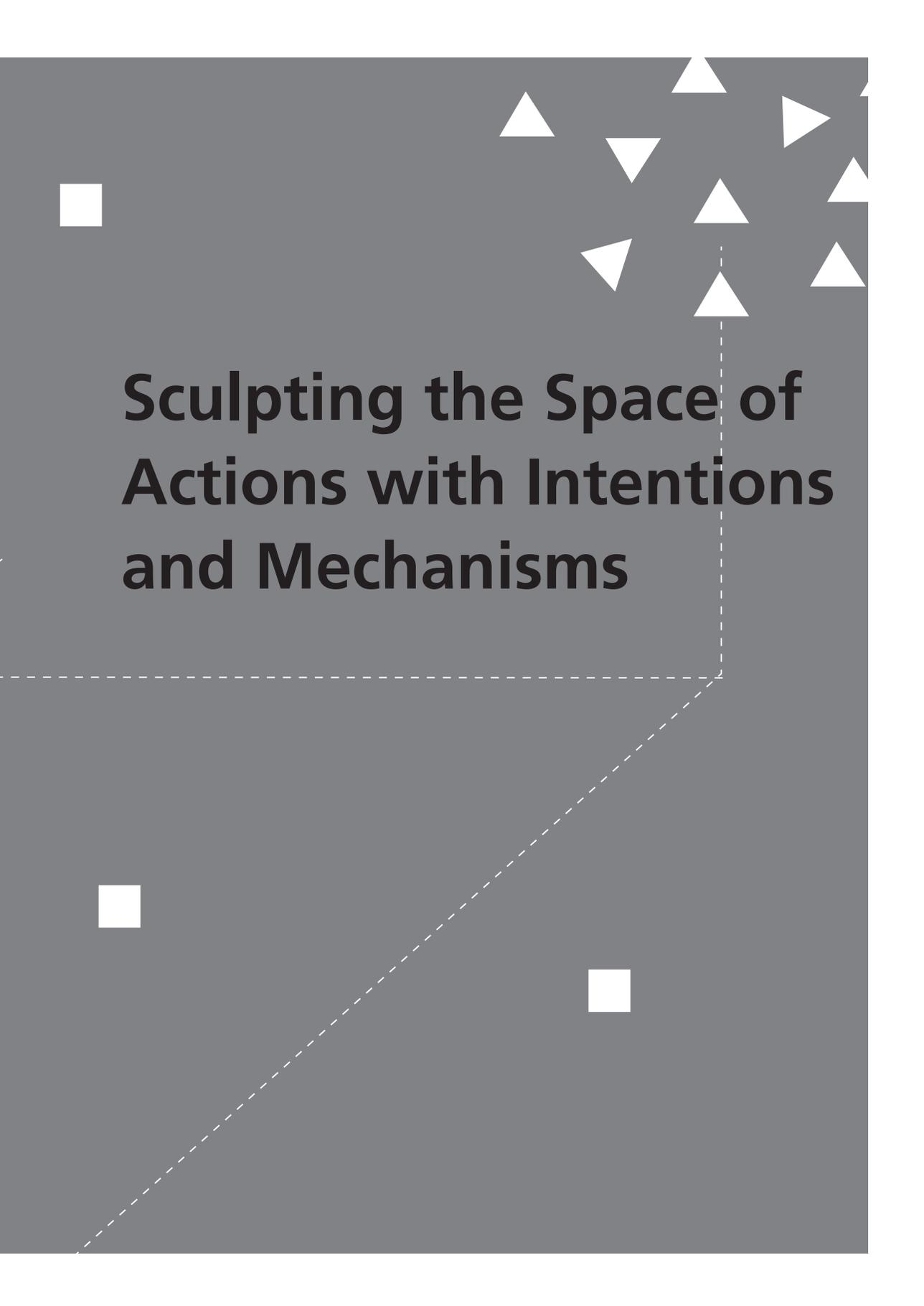
# Sculpting the Space of Actions with Intentions and Mechanisms

# 1 INTRODUCTION: MULTIPLE MECHANISMS YET STABLE PATTERNS*

The riddle which has been mentioned at the beginning of this dissertation and has guided us through the previous Parts will be put more in central focus in this Part. The riddle amounts to a paradox: to which actor do we ascribe more intentional control of his actions – to the expert who performs a complex action without the continuous, conscious selection and control of his actions, or to the novice who is almost incapable of performing that action as he has to continuously select and control all his movements and vocal sounds? On the face of it, one would perhaps ascribe more intentional control to the latter, but a second look offers good reasons for preferring to ascribe maximal intentional control to the expert. Nonetheless, perhaps some qualifying statements need to be added. But let us build up the question first.

The first steps of our argument did concern the kind of explanations that are available for cognitive functions, as they are underpinned by neural processes, influenced from without and within, and change over time due to development and learning. After considering different types of explanation, we argued that mechanistic explanation seems best capable of accommodating these properties of cognitive functions. Not only can we apply the three different theoretical perspectives articulated by Marr to all components that figure in such explanatory mechanisms, these are optimally prepared for accounting for the dynamical processes that cognitive functions are involved in. Learning to sing can be mechanistically explained as the recruitment of an additional component of tone control into the mechanism responsible for speech; steadfast voice control is then depending on the stability of the network that constitutes the mechanism; change in the mechanism's organization occurs when growing expertise is associated with automatization and the corresponding decrease in recruited network components; such a decrease does allow novel influences on the mechanism's activities, for instance when a singer's increased expertise enables him a flexible responsivity to perceived orchestral sounds.

The second Part offered insights in how a single function can be performed by multiple mechanisms. This is the more so, as mechanisms – complex and dynamic as they are - will always develop patterns of stability, partly as a result of a process of so-called kludge formation. Whether it is child development of voice control, or the automatization of certain associated patterns of stereotypical behavior, or the establishment of simulators that help to smoothly and comprehensively interpret and play a scene from an opera, or the seamless integration of external tools and objects

---

* On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively. Fig. III is particularly relevant as a representation of the main contents of section III.4.

in an actor's performance: the formation of kludges in the responsible mechanisms facilitate such developments. Due to the formation of these kludges, several additional characteristics could be added to the dynamical mechanisms involved in complex behavior. A few are particularly interesting for the topics of this Part.

First, even though we can observe the emergence of a kludge in the changing properties of the performance of a cognitive or behavioral function - particularly as the function is performed more coherently and consistently - we cannot derive from these properties the representations or information processes involved. Neither can we derive from the observable properties the neural implementation of the responsible kludge. Indeed, processes like development or automatization, described in the previous Part, may appear similar with respect to several properties even though the underlying representations may differ from case to case, as may their neural implementation.

Second, kludges emerge in the mechanisms not from scratch but usually by recruiting or re-using components that were already in place – corresponding to the modes of mechanism modification that were mentioned in Part I. Similarly, these kludges can themselves be involved in further developments and present building blocks for future changes in the mechanism. This is one of many phenomena that support the assumption that hierarchical structure is prevalent in many of such mechanisms and in the actions they perform. Indeed, it was mentioned several times that, generally, we can observe that early developments become so much involved in multiple later ones, that the early components become ever more generatively entrenched in the mechanism. For example, once a child's language production has become stabilized, the underlying mechanism will be entrenched ever more and can not easily be changed in an essential form. Consequently, a change in his language production will have a much larger effect on a person's overall performance than a change in a more superficial and recently developed capability.

A third and final – of the seven – characteristic to be mentioned here is the involvement of external information and even objects in kludge formation. Kludges are not only established due to the internal repetition of certain neural network activities, but also under the influence of external information in most cases. Whether it is the implication of culturally specific tone systems in early learning of vocal control, or the association of the manipulation of a sword or cross with difficult melodic lines, external information leaves eventually some trace on the kludge that is established during the learning process. As a result of this, we should recognize how an explanatory mechanism is partly characterizable in terms of the external information involved in its development.

In sum, until now we have more generally discussed the theoretical and empirical arguments that support the notion of cognition and behavior being produced by modifiable mechanisms that can establish kludges with the involvement of previously developed (or evolved) components and of external information and objects. Indeed, we have argued that the complex performances that an opera singer must give are made possible by this mechanism modifiability. Integrating speech and singing with acting and responding to other protagonists and the accompanying music, while flexibly adopting the interpretation of the director, an opera singer can do this only provided his brain is capable of developing mechanisms that include these kludges. So far, we have barely touched upon the question what role intentions play in all of this. On the contrary, it may have seemed that intention did not play a role earlier as we were mainly interested in phenomena like proceduralization and automatization.

Intentions will be put more central in this Part. To that end, we will discuss a differentiation between different types of intentions and scrutinize their interrelations. This discussion will be nourished both by philosophical analyses of action and the intentions that partly determine it and by empirical investigations of the processes involved. As we will find, the philosophical analyses are not only directed at the level of an agent's explicit formulation of intentions to act, but some also aim to clarify the contribution of component processes to his action – component processes that in themselves resist verbal articulation, like the motor intentions or the representations of motor movements (Pacherie 2008). We will apply and discuss the framework offered by Pacherie, which describes a 'cascade of intentions' as it offers an integration of three different types of intentions that together allows an analysis and explanation of human agency (Pacherie 2008). As these different types of intentions can all play some role in the processes that – sometimes after extended periods of time – lead to an action, the framework facilitates systematic discussion of these roles with reference to both philosophical and empirical insights pertaining to them.

Indeed, given the results of Part I we expect to find that a mechanistic explanation of action will indeed refer to several component processes that in an organized fashion, including interactions between these component processes, produce an action. Following up to the issues discussed in Part II, we will also investigate whether this framework allows room for the modifications that we found to take place during development and learning. Kludge formation, we did find, can have profound impact upon the mechanisms underlying the performances of an agent. Regarding the present context this raises the question whether such kludge formation has an impact on how an agent determines and configures his actions, affecting their complexity and temporal extendedness. What would it mean for this cascade of intentions when

specific actions have become automatized? Once his expertise has resulted in a large set of actions or a domain of action that have been acquired, practiced, and modified, will this have an impact on the intentions that he forms? Are these intentions thwarted by the prominence of these actions, which are often performed automatically, or is the converse the case: an experienced agent is much better equipped for intentional actions even if these do not always require his conscious control.

These and other questions will occupy us for the rest of this final part of our dissertation. We will follow the structure offered by the intentional cascade and amend it in certain respects. However, next to the distinctions it makes between distal, proximal and motor intentions, we will first explicitly determine the notion of a 'sculpting' process which leads to an agent's 'sculpted space of actions' into this framework. Furthermore, given our interest in the mechanistic explanatory approach, the question is whether the discrete distinctions suggested by categories of intentional action are correlated with different cognitive processes and perhaps also with corresponding different forms of neural activations? Moreover, we will consider whether there is a correspondence between the hierarchies that structure both the intentional cascade and the mechanism underlying it. This is the more interesting, as one could suggest that actions performed by an agent are different from a novice's regarding the kind of intentional control to which they are subjected.

Let us now first clarify what we mean with a 'sculpted space of actions' before presenting the intentional cascade at more length. Together these two discussions will allow us to formulate some features of the framework that we consider necessary for the explanation of an agent's increasingly complex performances.

## 1.1 'Sculpting the space of actions' – an important ingredient for the explanation of expert action

Human action allows no simple explanation, as a causal pluralism is involved in its production and we argued in Part I that a corresponding theoretical pluralism is required for its explanation. We rejected a Socratic approach to intentional voluntary action, that holds that it is rationally deduced from an absolute moral principle. Instead, we embraced the Aristotelian account with its causal pluralism, according to which, for example, moral principles can somehow become internalized in an agent's habits and dispositions, implying that these principles can exert their influence via more than just a single representational format and a single process. Apparently, moral deliberation can yield results that are somehow accessible and useable by psychological processes as different from such deliberation as habits and dispositions. In sum, different as these processes and the contents involved may be, they interact and influence each other

and together result in an action. Modern studies have confirmed the causal pluralism behind action. Neurophysiological processes, affective motivations, memories and expectations, rational deliberations, environmental conditions, social influences and many other factors interact with each other during the performance of an action. In addition, development and learning create differences between novice and expert action in many ways, as we learned in Part II. Representations and mechanisms change and influence the properties of an agent's cognitive and behavioral responses, adding complexity to the pluralism. Yet somehow all of these factors converge and together produce, or result in, an action. So how should we conceive of this complex and dynamic process, given the various responsible component mechanisms involved in it and the various types of informations and their transformations that are being considered by the agent?

We propose to view this problem of determining an action by an agent as a search for a suitable candidate action in a multidimensional space – a space of action options. This action space is influenced by a multitude of factors, both dependent upon the agent himself and upon external factors. The action space's shape will be modified in a relatively stable way due to development and learning, yet will remain adaptive as it also responds to ongoing internal and external conditions.

To offer a first explanation of this view, let us consider a similar framework applied to a language processing task. Take, for example, the fact that we are capable of speaking fluently and use thousands of words while doing so. Now finding a word to begin or continue a sentence can also be considered as a search problem: a problem to find one or more suitable options in a large space of options. Particularly for those with expertise in a language, there are usually many alternatives for each word, even more when they have mastered several languages. These alternatives are not identical, though, differing in terms of semantics, grammatical and syntactical properties, idiomatic meanings, associations, and so on. Each of those factors can function as a constraint on the space of options available for an appropriate word that is to be expressed, constraining the search problem somewhat and alleviating the task accordingly. For example, when we write English sentences, the space with suitable options is restricted, as Dutch, German, Greek and other words are excluded from it. At times, a multilingual speaker will inadvertently insert a word from another language into his English speech – the surprise and annoyance about this signals the fact that it is quite exceptional. Similarly, once we have chosen to use the *pluralis modestiae* or *pluralis auctoris*, words like "I" and "me" and the like are excluded – though these may appear in quotes. In a dissertation, expressions like "I just believe that…" or "It is stupid not to understand" are to be avoided even more than grammatically incorrect sentences. In short, there are many

constraints at work at several levels of specificity and different in kind, that can limit the search problem somewhat – if, that is, the author's grasp of the language and the conventions of the trade is appropriate.

Studies with word generation tasks by Chris Frith and others confirm the notion that finding a word in a space of options is influenced by constraints upon that space. Completing a sentence or filling in a blank in a sentence has been shown to involve various component tasks like generating words, selecting from a particular set of words, checking words for different sorts of appropriateness, inhibiting inappropriate options, and so on. Associated with these tasks, multiple component mechanisms have been identified, along with multiple representations and transformation of contents at several levels of specificity. Frith explains the fact that dorsolateral prefrontal cortex (DLPFC) activation increases in cases where there are only few constraints on the space of options provided by a particular sentence. Here, the subject somehow has to determine by himself (or herself) a selection of options and pick a final answer, and DLPFC appears to be involved. Frith describes this as: "the "sculpting" of the response space normally achieved by external context that has to be self-generated" (Frith 2000 560).

'Sculpting the response space' is understood as a dynamic process that is determined by factors that can be internal or external to the subject, stable or dynamic. Internal factors that influence this sculpting process are the language expertise of the subject but also cognitive and neural factors like his memory and the current stress hormone levels. External factors matter, as when the subject is confronted with easy or difficult task sheets, but the treatment by the research assistant also influences the process via the stress responses that it provokes. Some of these factors are relatively stable, while others dynamically influence the process, for example a given situation or even a particularly shocking word in a particular sentence. Each factor refers to a different (component) mechanism that will involve a specific representation and transformation of information. Nonetheless, as finally a particular response must be given, all factors must somehow converge in the process of determining a single option from the response space – each in its own way.

Such a response selection from a large space of options is relevant in many domains other than language. Indeed, this process of 'sculpting the response space' is held by Frith and others to obtain in different modalities and task domains, including action selection (Frith 2000).[279] They have devoted some research to further determining the set of component tasks and elucidating the impact of particular constraints on this modification process (Fletcher, Shallice et al. 2000). For example, a study with different representations of information and their transformation has shown how these

influence the process, too. Rule-based selection turns out to be important, for example, recruiting particularly DLPFC activation, while other component tasks involve other forms of association and are carried out by other neural activities (Nathaniel-James and Frith 2002).[280] Indeed, studies like these inform us about the mechanisms and respresentations involved in particular instances of sculpting a response space, shedding light on the specific properties of the space itself: its dimensions, its flexibility, its structure, and so on.

As attractive as this framework of a search problem being influenced by a sculpting process seems to us, it has been applied only to a limited extent in the domain of human cognition and action but perhaps more in the domain of AI and robotics.[281] Let us consider a few prominent examples of such applications before arriving at our notion of 'sculpting the space of actions'. The explanation of human color perception may be the domain in which a spatial or geometrical framework is most widely used. For example, colors have been taken to be points in a multidimensional space determined by their various properties, like value, hue and chroma (Munsell 1912). Meanwhile, certain properties of human color perception are explained with reference to a spatial color representation in the form of a spindle.[282] In this way, a range of phenomena can be explained, like the perception of after-images, the effects of contrast colors and the differences between languages in how they carve up this space with their respective color vocabularies (Regier, Kay et al. 2007).

---

[267] Another experiment in which subjects were asked to engage in 'willed action' by lifting at random their fingers showed DLPFC activation patterns comparable to those that occurred in word generation tasks. This concurred with observations in other experiments that required subjects to act or move (Nathaniel-James and Frith 2002). Indeed, there is growing consensus with regard to the overlap in neural activations for language processing and action processing, see for example (Grèzes and Decety 2001 ; Pulvermüller 2012; Pulvermüller, Hauk et al. 2005 ; Raposo, Moss et al. 2009 ; Taylor and Zwaan 2009 ; Willems 2009). This suggests that the representation of the content of speech and action do overlap to a large extent. An implication is that such representations are available for several processes, offering a crucial role for simulation as a prevalent form of computation (Barsalou 1999c ; Jeannerod 2006).

[268] Interestingly, in conditions when the response space is less constrained – by contextual clues, for example -, subjects show more DLPFC activation and take longer to respond (Frith 2000).

[269] Supporting the usefulness of multidimensional spaces as the representational format of information involved in cognition and action is its feasibility for the construction of simulation and robot models (Gärdenfors 2004b ; Gärdenfors and Williams 2003). In a robot, for example, decision making can take place at the level of conceptual spaces, where all relevant information is represented. Here, perceptual constraints determine a set of possible actions that are preselected on the basis of the explicit programmed instructions (Chella, Gaglio et al. 2001). Artificial decision procedures can be developed in this vein as well. For example, expert systems in clinical situations make heavy use of 'relative magnitudes' pertaining to specific dimensions and which stem from counts, measures, weights, and so on. Gärdenfors' framework as a 'meso level representation' is considered helpful for such construction work (Aisbett and Gibbon 2001). In all cases, however, the number of dimensions and their relations have been decided upon by an engineer and have not evolved and developed naturally, nor are they allowed to be modified in unpredictable ways as is the case in human agents due to their development, learning and experience.

Another example concerns taste. It is argued that particular tastes are the result of specific activation patterns elicited by the four taste receptors and consequently occupy specific areas in a multidimensional 'taste space': "[i]n this way are the brain's representations of the various possible tastes arranged in a systematic "space" of similarities and differences" (Churchland 1995).[283] Although these taste perceptors are activated by different electrochemical processes, their activations are represented in a common space and arranged together in such a way that they give rise to our recognition of thousands of different tastes.

In a similar vein, Churchland has applied this framework to explain how an agent determines his action in a world full of objects. Instead of separating the processes and representations involved in his perception of his environment, his decisions to move and the performance of the movement, Churchland argues that these processes should be taken as employing a shared multidimensional space. Evidence for this comes from studies demonstrating the interactions and interferences between those processes, among others. This implies that the spatial representation of object location and the spatial representation of bodily movement are somehow integrated in the same representation space instead of employing two separate spaces. If the latter were the case, it would require a complex translation process of his spatial movements for placing them in a space that represents the environment, if the agent wants to avoid bumping continuously into external objects. Indeed, the integration of both representations into a single space would facilitate his sensorimotor coordination and thus enable the agent to: "assume[…] a position in its "motor space" that corresponds to the position of an object in its "sensory space" (Churchland 1995).[284] Evidence suggests that this integration is indeed the case.

---

[270] It is debated whether color vision or taste are more dynamic processes than is often thought, affecting their representations too. Investigation of neural firing rates related to taste perception in rats under different conditions, demonstrates that taste perception is a highly dynamic process, modulated by other cognitive processes or states of the animal. As a result, hedonic impact and incentive salience of a taste are variably modulated (Tindell, Smith et al. 2009). Modelling such changes via adjustments of a state space would require continuous modifications of the geometry and topology of that space, leaving behind some of the attractive simplicity of such representation.

[271] The spatial arrangement of taste representations is also used by Churchland to distinguish between tastes that are prototypical for a particular fruit, for example, hyperbolical divergences of these, and so on. In addition, with such a spatial format of representation, relying as it does upon the number of dimensions combined with the levels of discrimination within each dimension, we can also determine the size of a state space pertaining to a particular cognitive function and compare it with a space of other functions or with a similar space in other animals. The state space of smell, for example, is much larger and contains much more levels of discrimination in dogs than in humans (Churchland 1995).

[272] Compare this issue with the theory that sensorimotor coordination is enabled by 'common event codes', that is, by codes of features of perception and action plans stored in a common representational medium (Hommel, Musseler et al. 2001). This theory is modest in its domain of application and specific in that event codes are presumably shared by both tasks and how they are implemented.

It is important to realize that such a multidimensional space is employed here as an explanatory tool. It offers a plausible representation of how the results of different cognitive processes appear to be related to each other. In other words, it is a second order representation, a representation of the results of multiple cognitive processes. Such a second order representation is different from the two representational formats that are prevalent in cognitive neuroscientific explanations: the representational format of symbols and propositions and the representational format of connections among neurons.[285] An important characteristic of the multidimensional, spatial representation is that a fundamental role is played by similarity-dissimilarity relations. In this way, we can explain how color perception appears to employ a continuous color-space, while this space simultaneously appears to be carved up into separate sub-spaces with verbal categorization.[286] Such an explanation of color concepts within the same framework as color percepts is parsimonous, indeed (Gärdenfors 2004b 2). This framework has been used to explain other features of cognitive processes as well.

Important for our purposes is whether dynamic factors like someone's expertise or an environmental condition are allowed in this framework, changing for example the contents and shape of his or her representational space. Gärdenfors indeed argues that learning or development (and even evolution)[287] can be explained in terms of the change of contents or structure of a relevant space.[288] Such changes in a person's representational space for a particular domain can be stable but there are also dynamic

---

[273] Gärdenfors concurs with the critique of the sentential conception and the defense of a spatial account of information representation given by both Churchlands (Gärdenfors 1996).

[274] A category is a particular region of a conceptual space that has been carved up. Even though the conceptual space is itself continuous, it can be carved up into regions with sharp boundaries. For example, even though the color space is continuous, color terms suggest sharp boundaries between colors (Gärdenfors 2004b). Probably due to visual physiology there is large agreement between languages in how their color terms refer to regions in the color space (Kay, Berlin et al. 1991). Nonetheless, there are differences between separate languages in the number of color terms they use or the boundaries drawn between color terms. Interestingly, experiments with English and Korean subjects shows that 'categorial percpetion' occurs as subjects' perceptual discrimination corresponds with the category boundaries of their language, suggesting an important role for 'categorical perception' (Roberson, Hanley et al. 2009).

[275] Evolutionary processes have also effectively contributed to the determination of spaces for several functions. Gärdenfors argues that one could easily reformulate Marr and Nishihara's explanation of visual object shape recognition in terms of the employment of a 'shape space'. According to them, evolution seems to have resulted in a rather simple process depending upon the fact that biological objects tend to have a form based upon generalized cones. Consequently, for visual recognition only a limited set of dimensions needs to be processed, like the size of the cones that are connected to each other, their orientation, the components' axes and their reciprocal configuration (Marr and Nishihara 1978). This approach therefore allows the representation of biological objects within a multidimensional space of options.

[276] The author defends the representation of Piaget's findings of how children learn to differentiate between height and volume when perceiving filled glasses according to his geometrical account of cognitive processes (Gärdenfors 2004b). To be sure, this development can also be represented differently, for example as an event along the lines of catastrophe theory (Molenaar 2001).

changes of his or her representational space that obtain as a result of environmental or internal processes. For example, a process like the direction of attention to a particular property can be considered as affecting the agent's representational space. When he is attending specifically to an object's weight he can make finer discriminations, which can be represented as 'stretching' the distances along that dimension. Conversely, neglecting such a property amounts to 'shrinking' these distances (Gärdenfors 2004b 20).[289] Many more internal and external can contribute to such changes of an agent's representational space, affecting among others the response space that his cognitive or behavioral response depends on.

The discrimination of actions has been described in terms of such changes of a representational space, brought about by a possible neural network which learns to distinguish two dimensions. More specifically, within this spatial framework it is explained how the network might represent and discriminate between moral dimensions of an action. By judging the similarity and dissimilarity between actions along some relevant moral dimensions, the network would place actions like assisting, murdering, lying and self-sacrifice at various locations in an action space.
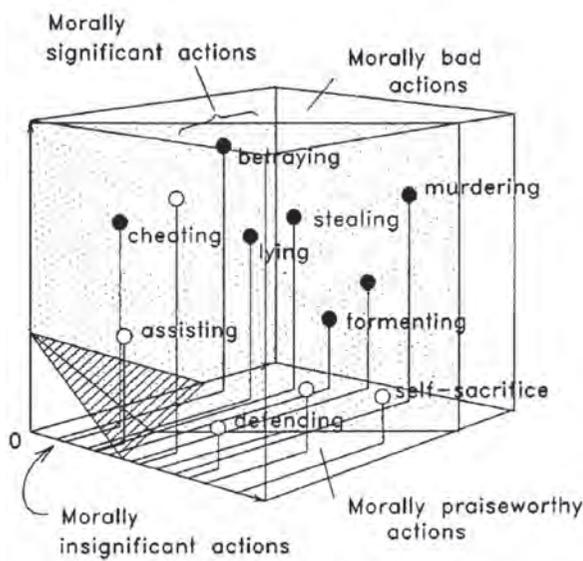


**Figure 2. A (conjectural) activation space for moral discrimination**
Reprinted from (Churchland 1998) with permission from the publisher.

---

[289] In Churchland's terms, such refined discrimatory ability would amount to an increase in the levels of discrimination with respect to a certain state space or one of its axes (Churchland 1995)

Paul Churchland and others argue that an agent does not usually determine his actions via deduction from abstract moral principles but by employing such a framework of action representations in an action space. That is not to deny that irrespective of the spatial contiguity of this space, moral concepts and judgments tend to carve it up rather strictly, as do color concepts with the multidimensional color space (Churchland 1998; Casebeer and Churchland 2003). In figure 2 above, this is represented by the vertical pane that separates praiseworthy from bad actions and by the diagonal pane that leaves a corner for morally insignificant actions.

What is not visible in this framework for the spatial representation of moral actions is whether its representations are flexible as a function of the conditions under which it is employed. For example, we would maintain that for most agents, no action is morally insignificant under all circumstances – and perhaps vice versa. Moreover, when an agent is about to determine an action, the dimensions of his space of actions will flexibly respond to internal and external conditions like his emotional state or the risks provided by the environment. These conditions would have an impact on the dimensions and the structure of the spatial representation and with that, also on the placement of the actions in it. It may even be the case that the conceptual distinctions he usually makes – represented as panes in the figure – will shift or that an individual action will change sides.

Let us now apply what we have learned from this short discussion to the notions that we are introducing in this dissertation: sculpting the space of actions, and a sculpted space of actions. Whereas Frith's framework of 'sculpting the response space' primarily referred to an ongoing dynamic process of determining an appropriate answer for a given problem, the other spatial representations rather contained stable representations of a domain like color or moral action as used by an agent. Frith's framework implied that internal and external constraints help to constrain the response space in a particular case and facilitate a final response choice. The other frameworks emphasized how the stable representation of a domain can change due to development and learning in terms of its size, its dimensions, its structure and so on. Our framework, finally, has the ambition to combine these two, realizing that dynamic and stable properties of a sculpted space can and should not be separated but integrated.

Indeed, we noted that according to Frith, the process of 'sculpting the response space' was also constrained by the relatively stable properties of the space due to the agent's expertise. These stable properties are themselves the result of a long-term sculpting process, since development, learning and practice will affect an agent's representational space pertaining to a particular domain. The more this has resulted in

a stably 'sculpted space', the easier it usually should be for an agent to find an appropriate response.[278] In other words, finding an answer to a particular question amounts to dynamically applying *further* constraints to a representational (sub-)space that is itself already sculpted or constrained as a result of expertise. For example, the recognition of relevant constraints in a given situation depends upon the agent's mastery of the domain, such as by his knowing musical structures and rhythms.

Yet this interaction between the stable and the dynamic properties of a sculpting process is itself complex.[279] Irrespective of the agent's mastery, this recognition of relevant constraints can easily be impeded by other cognitive and neural processes that simultaneously take place. Attention, for example, was already mentioned to potentially affect a dimension of the representational space employed by the agent, stretching or shrinking it. Stress was mentioned as well, as a dynamic factor that can influence several processes involved in the sculpting process. However, in the case of an expert, we expect him to be less vulnerable to such disturbing factors and to perform more reliably and appropriately than a novice, attacking the correct note at the appropriate time, for example. In other words, the stable space that he has established over a long time of sculpting has such properties that it allows him to determine an appropriate answer even in cases where he is affected by distraction or stress. So, in an expert singer's sculpted space of actions the tritone or *diabolus in musica* occupies a very small place in a far corner of the space, making it unlikely for him to sing it even when he is tired. Moreover, his expertise should also imply that he realizes how such conditions dynamically affect the sculpting process, knowing strategies for countering them. Being tired, he pays extra attention to what the orchestra plays and

---

[290] It might be maintained that there are cases in which an expert, having a sculpted response space, can come up with a response that is less adequate and innovative than a novice's. Such a response would be represented in a relatively small and peripheral subspace in the expert's action space. Such a case might indeed occur, yet it is likely to be a chance hit because what is characteristic of a novice is his limited expertise with regard to the constraints and properties of the domain. We concur with Boden, who argues that one should not consider such a novice's response to be genuinely creative (Boden 2004).

[291] These spatial accounts also emphasize the temporal dimension of cognitive processes, often by integrating some form of dynamical systems theory. This is offered as being superior with respect to its handling the temporal dimension in cognitive phenomena than traditional approaches like the traditional computational approach (Van Gelder and Port 1995). Such a dynamical systems approach can also be used without talk of representations, even though the latter is common in cognitive science. Instead of content being represented in a distinct way within a computational system, different vectors – for example, one for every taste receptor – determine the state that a particular system is in. In that case, the state of the dynamical system at any moment carries the necessary information without there being some sort of representation present in the system (van Gelder 1998). One could argue that this is not more than a semantic difference between the approaches, differing as they do in their definition of what a representation is. Furthermore, there are many cognitive functions in which correctness or incorrectness of the representations involved do matter. Dispensing with representations altogether would therefore not be advisable (Bechtel 1998).

to the preparation of his voice in order to sing the right note.

Finally, it is important to remark that this representation of actions in a multidimensional space of actions is not at odds with further references to other action representations. The multidimensional space allows us to add further properties or dimensions to the actions that are represented in it, as long as each dimension can be processed by a cognitive or behavioral function. For example, we will next discuss how an agent's actions are influenced by the different kinds of action intentions that he has, differing in their format of representation, among other things. Motor intentions, for example, are in the non-conceptual format of sensori-motor representations, whereas distal intentions are verbally formulated and in a conceptual format. Expertise with these different levels of intention regarding a particular action will affect the action's place in the agent's space of actions. For example, it may be that a novice has not yet practiced the sensori-motor representations that belong to a certain action, making it less likely that the action will be selected from his action space in an emergency. When an expert has extensive experience with that action and he is required to act, this action associated with sensori-motor representations will figure more prominently in his constrained space of actions. Expert action, then, is indeed dependent upon sculpting the space of actions along the lines described here. The sculpting process involves a combination of both expanding and constraining the space of actions. Important to repeat is the fact that this sculpting process is both a long-term and a short-term process: it contributes both to an agent's stably sculpted space and to the dynamic properties of his responding in a given situation. Without such sculpting the space of actions, it is difficult to see how expert action can be performed at all.

## 1.2  Determining an action via a cascade of intentions

An expert singer may have noted that his recent performance of Don Giovanni had unwillingly been affected by his solemn interpretation of Saint François not long ago. Intending to correct this flaw in his interpretation by making his Don Giovanni more boastful in general, he realizes that some solemnity may still work in the dialogue with Donna Elvira, who is naively trying to convert her unfaithful lover. Therefore, even after having formulated a general intention for a more boastful Don Giovanni, our singer needs to more specifically remain alert for situations in which this intention can prudently be acted out, while avoiding others. Needless to say, that such intentions and performances can only usefully be made by a singer who has such a mastery over his singing that he can switch timbres at wish from solemnity to boastfulness (and change his style acting correspondingly).[292] This description of a not uncommon form of self-correction and self-control involves three different types of intentions,

which we will introduce in this section. We can distinguish between these types of intentions in terms of their contents, their functional roles, their temporal trajectory, and so on. Important, however, is the fact that they are intimately related to each other, contributing to his coherent and consistent performances of different roles. Let us consider a model that aims to account for this impressive feat.

Integrating philosophical and empirical insights, a model has been developed to account for the hierarchical control of motor action via a 'cascade of intentions' that spans different levels of specificity of an intended action in (Pacherie 2006 ; Pacherie 2008).[293] This model considers intentional action in its most comprehensive form as the result of a process that can be described as a process with three discernable and distinct phases, starting with a deliberate intention to realize a future goal and completing when particular muscular activities have realized that intention.[294] Each phase has a different functional role, involving different formats of representation and transformation. These phases do not strictly succeed each other, with an intentional action possibly occurring without contributions of all phases of the complete cascade. In closing this short description of the framework, it should be noted that there are many different forms of interaction between the phases.

The model distinguishes between three different forms of intention, to wit: distal intention, proximal intention and motor intention. Distal and proximal intentions have been borrowed from several philosophical accounts of intentional action, while motor intentions were added to those on the basis of evidence from the cognitive

[280] Research of expert performance in domains as far apart as sports, music and science has demonstrated that extensive periods of deliberate practice generate the necessary cognitive and physical adaptations for exceptional performance. Improved motor performance also requires enhanced cognitive representations and skills, contributing to its improved selection, guiding and correction of motor actions (Ericsson, Roring et al. 2007).

[281] It must not be left unnoticed that this three-level model of intentional action has been used particularly for the explanation of the phenomenology of agency. To that end, the model has been further equipped with feed-back and feed-forward relations between levels of intentions and comparators that serve to discern congruity or discongruity between intentions, motor movements, perceptions (Pacherie 2008). Apart from such specific use, however, it can still serve for the explanation of the 'generation and control of action' (Pacherie 2006). The model allows still further elaboration or expansion, for example with the integration of the What, When, Whether model of intentional action, as the authors of both models have hinted at in (Pacherie and Haggard 2010).

[282] There are several others models of intentional action available, of course. Depending on their explanatory or analytical focus, these models differ from each other. For example, the WWW model of intentional does not so much focus upon the different forms of intentions involved, but rather on the different component decisions involved in intentional action, that is: the decisions about what, when, and whether to do (Brass and Haggard 2008). A different model of intentional action does also integrate cognitive – decision making - and motor processes like Pacherie's model does (Cisek 2006). These authors use the notion of 'representation' as a common denominator for all processes involved, eschewing the notion of 'intention'.

neuroscience of motor action control.[295] Several philosophers of action have argued that it is important to distinguish between distal and proximal intentions since these play different roles in the various phases and features of actions.[296] Most prominent, of course, is the difference in their being oriented towards future actions or being aimed at the realization of an action here and now in the case of proximal intentions. Relevant to note is that there is a reciprocal interdependency between the different temporal orientations of these two forms of intentions: if a proximal intention is carried out without any orientation upon the agent's distal intentions, the agent runs the risk of frustrating and even counteracting his own long-term intentions or might fail in the coordination with some other intentions of himself or interested other parties. On the other hand, realizing a distal intention requires the recognition of a suitable situation and anchoring the appropriate action in that particular situation, which is the role of a proximal intention. The performance of that appropriate action in a particular situation, which realizes the distal intention, finally relies upon specification of the necessary muscular movements that are captured by the motor intention.[297] Given that such motor specifications, while necessary for flexible performance of intentional action, escape the kind of awareness and explicit control that can be applied to the other intentions, these differ sufficiently from those other two forms of intentions to merit separate mention.[298]

---

[283] Pacherie (Pacherie 2008) makes reference to Searle's distinction between prior intentions and intention-in-action ((Searle 1983); see further below for its relevance in guided action), Bratman's distinction between future-directed and present-directed intentions (Bratman 1999a) and finally Mele's distinction between distal and proximal intention which was adopted by her (Mele 1992). Pacherie notes that it was mainly the absence of temporal connotations in the latter distinction that made her prefer it above the others. Our discussion, further below, will stress the importance of avoiding incoherence and inconsistency between actions, which is more at the center of Bratman's arguments.

[284] Actually, Pacherie distinguishes even seven different functions of intentions: intentions can function both as prompters and as terminators of practical reasoning; they serve individual and social coordination; they can function as initiators of a performed action and serve to sustain an action until the end; performing meanwhile a guiding function while also assisting in monitoring the adequacy of the action's performance (Pacherie 2006).

[285] Contrary to common use of what 'intentions' are taken to be, these motor intentions contain neither propositional content nor are agents usually aware of these in this framework (Pacherie 2008).

[286] Within the modern Anglosaxon domain of philosophy of action, it may have been Frankfurt who was the first to point out the relevance of such automatic motor adjustments for intentional action. Moreover, he has also pointed out that the resulting 'purposiveness of our behavior' it not limited to human action as also spiders must be said to act intentionally, for example – albeit in a weaker sense than humans do (Frankfurt 1978). In continental phenomenology these topics had been debated much earlier, particularly by Husserl, Merleau-Ponty and some oher phenomenological authors (Painter and Lotz 2007). This tradition has been inspired by Aristotle's philosophy of biology in which the animal's responsiveness to its environment figures prominently (Oele 2007). This is another example of a long forgotten lesson from Aristotle's biological works that has impeded philosophy, as it emphasizes continuity and gradual differences between different animal species and mankind instead of focusing mainly on distinctions and divisions.

Indeed, there is a dynamical interaction between all three forms of intentions. In fact, the interactions at stake can flow in two directions. To be expected is the top-down flow of control in which a distal intention refers to an agent's particular long-term goal, waiting for the appropriate situation to present itself for fulfilling such an intention by anchoring it in that situation by determining the corresponding proximal intention. Subsequent performance of the latter is dependent upon the motor intentions that specify the necessary movements. In addition, a bottom-up flow of control must be acknowledged, both on phenomenological and neuroscientifical grounds. For example, the mere perception of an environmental object that can be involved in a motor action does provoke activation of the corresponding motor representations, without an agent's deliberate intention to act (Grèzes and Decety 2002). Indeed, the relative autonomy and independence of this lower level of action control is such that: "the affordances of an object or situation are automatically detected even in the absence of any intention to act" (Pacherie 2008 186).[299] Subsequently, upon these motor representation activations, corresponding proximal intentions may arise, as when an agent may realize only his desire to quench his thirst when he finds himself reaching unwittingly to a perceived cup. Finally, an interaction between levels of intention may also occur in order to correct or interrupt an action, for example if motor movements must be adjusted or even interrupted due to a changing environment. More below in this Part we will further discuss such dynamical interactions.

Summing up the foregoing, we can refer to a 'cascade of of intentions' that together comprise a hierarchical model of action control as is visible in Figure 3 below from (Pacherie 2008). In this figure, along the vertical axis, we see how P-roximal and M-otor intentions are subsumed under D-istal intentions. Horizontally, we can discern how it may take a while for an intention (dotted line) before it enters into the process of undergoing the necessary transformation via situational anchoring or parameter specification (downward along vertical lines). As a result, an overt movement occurs. Conversely, as mentioned above, a bottom-up form of control can happen when in the absence of proximal and distal intentions, for some other reason an overt movement is made, triggering situational guidance and control such that the associated proximal goal is reached. At times, however, this movement may be interrupted or inhibited when the movements take long enough for an agent to exert an additional form of (rational) guidance and control, particularly when he becomes

---

[299] The clinical syndrome of utilization behavior suggests that indeed automatic detection of a potential opportunity for motor action can then lead to an action without the agent having a proximal intention for this action. In such cases, an agent may put up a pair of glasses even though he does not need a second pair upon his nose (Sumner and Husain 2008).

aware of its contradicting his distal intentions. As all these processes have their own temporal constraints, it is among other things dependent upon the tempo of the events whether the complete cascade of intentions can unfold, or not.

Having presented this model, it is relevant to underline again the use of putting these three types of intentions that have been distinguished philosophically into a single dynamical model, even though several differences between them have been noticed. Remember that in our Part I, we noticed that for the explanation of a particular cognitive or behavioral function we can develop an explanatory mechanism that captures the function at several levels of mechanism. Obviously, all intentions are the result of some cognitive processes going on in the brain and their realization does equally require an execution that at least involves some motor processes. Correspondingly, we can take the philosophical analysis as a heuristic and investigate whether its ingredients allow to be integrated as components in a more comprehensive explanatory mechanism. This will by no means be an easy task, as there are many
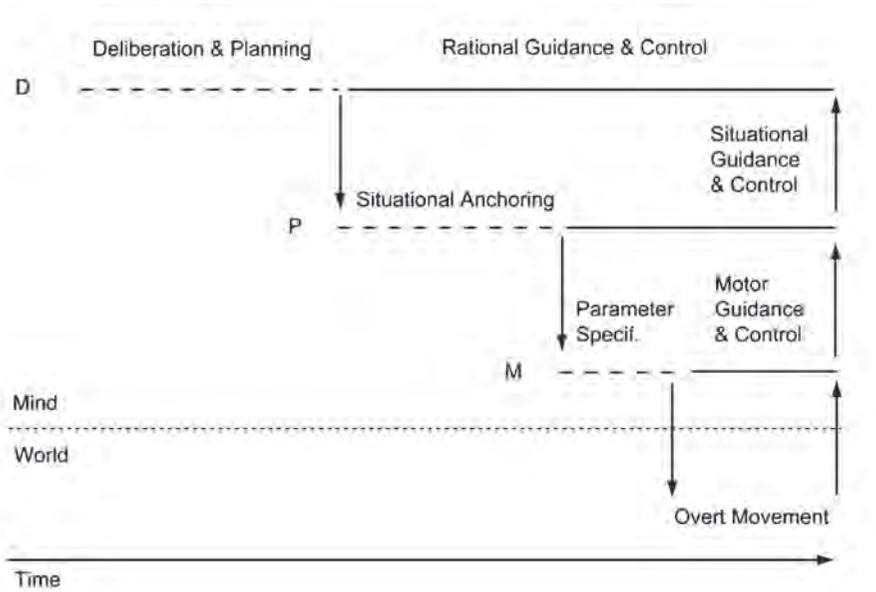


**Figure 3. The intentional cascade of D(istal) intentions, P(roximal) intentions, and M(otor) intentions.** Note that a horizontal *dotted* line refer to an existing intention still waiting for the phase of its further realization. A horizontal *continuous* line refers to the phase in which an intention is actually realized. Adapted from (Pacherie 2008 189) with permission from the publisher.

feedback and feedforward influences between levels involved, processes taking place at different time scales,[300] and so on. Indeed, if an agent is to perform coherently and consistently, such influences and processes must be connected to each other. Therefore, instead of keeping a philosophical analysis of action intention separate from a cognitive scientific explanation of motor action, it is a challenge to see whether a mechanistic explanation allows us to integrate these. Indeed, such an integration would invite us to note that the 'what' or the goal of an action "can be specified at the three levels of M-intentions, P-intentions, and D-intentions" (Pacherie 2008 196). Similarly, the model would allow a specification of the 'how' or the means of an action at several levels of specificity, as it does of other action features.

With this model in place, we can analyze how the performance of an action is produced by a mechanism that consists of different interacting sub-processes, which can in turn be analyzed from different disciplinary perspectives – including a philosophical perspective. The model offers us also a framework to further explore some of the issues that we found to be relevant in explanation of cognitive functions. The first issue is the algorithmic theory of the nature of the representation of information. As the model suggests and Pacherie has also noted explicitly – see the quote in the preceding paragraph - , we can expect different representations at the three levels of intentions: verbal in the case of distal intentions and in the form of non-verbal motor representations in the case of motor intentions. It is particularly interesting to consider what representations are involved in the intermediate level of proximal intentions and we may expect an interesting confrontation between philosophical analysis and empirical insights in that context, given its position between those explicitly verbal distal and non-verbal motor intentions.[301] Since we've learnt in the first Part that it is impossible to directly derive this representational level from either the task level or the neural implementation level involved, investigating the actual form of representation is challenging.

The second issue which particularly interests us in light of our investigations in Part two, is how we can integrate in this model a central role for experience and action skills.

---

[300] Indeed, apart from a control hierarchy that is responsible for the increasing specification of actions, it is important to acknowledge that actions also require a hierarchy of temporal extension as all actions are temporally extended. Both hierarchies do not necessarily overlap and require to some extent different neural and cognitive resources (Uithol, van Rooij et al. 2012).

[301] Pacherie refers to one of the two visual streams, to wit the 'vision for action system' in this context. This system allegedly produces motor representations in an appropriate format, usually involving an objects as an action goal, while taking into account several biomechanical constraints. The motor representations are then used in two different forms, as preditive or forward models and as inverse models (Pacherie 2006). The picture that emerges is still very much top down and still suggests a rather one-to-one correspondence between the distal and proximal intentions and their motor counterparts.

Given our findings regarding the gradual and dynamical processes of automatization and habituation of intentional actions and the corresponding differences between a novice and an expert singer, these processes could further demand elabortion of this model. As a result of the formation of several kludges and other changes through experience, we can observe in an expert how he has built up a space of actions that appears in many regards different from the space of actions the novice draws from. An important difference is that an expert can draw from this space without having to explicitly determine a particular action in detail, but switch completely his singing and acting from a solemn to a boastful Don Giovanni in an instant. Indeed, this space of actions can be 'sculpted' according to a particular dimension of the actions currently deemed important, to which then increased attention is directed. In such cases, a complex interaction takes place between an agent's previously sculpted space of actions and his intentional cascade – a phenomenon that we will certainly consider more closely in this Part.

In sum, this model is interesting in that it suggests how we can combine both a philosophical account of different types of intentions with empirical insights in the determination or control of motor actions. At the same time, it invites us to further explore two issues that we earlier found to be relevant for explanation of cognitive functions: the representations involved in these and the role of experience and skill in their performance. In line with the latter issue, we will also be alert to find whether we can observe in the literature reports of the same benefits that we previously found to adhere to hierarchically structured complex mechanisms once they dynamically develop and change, for example by sculpting a space of actions. Observation of an expert opera singer suggests that he indeed has the advantage of such benefits, as does the audience that must not fear for being disturbed by any instability of his voice, by his incapability of harmonizing with the orchestra and other actors, nor by his forgetting his acting once the singing becomes difficult.

## 1.3 The cascade of intentions and a sculpted space of actions

What we are interested in this Part is the complex process that leads up to an action which includes the involvement of different representations of action in and the role of a sculpted space of actions therein – which probably also adds to the multiplicity of action representations involved. At first sight, a philosophical analytical approach is most appropriate for analyzing the explicit and verbal or symbolic representation of action, as it is found at the level of distal intentions in particular. However, the model of the intentional cascade discussed in the previous section has been developed by combining further philosophical insights in the process and structure of action

with scientific insights in these, adding two other formats of representation of action. Our aim is to further add to this framework in order to develop it such that it can also account for the aristotelian observation that even moral action can over time be habituated or automatized and still not loose its moral significance.[302]

Therefore, we need a comprehensive framework that allows us to explain how an agent is capable of sculpting and employing a space of actions, partly determined by intentions and ideals to which he has committed himself earlier. Such a space of actions will allow him to act flexibly in a fast, stable yet also implicit manner, while still respecting some important constraints for action that he has taken upon himself. So we are interested in a framework that can explain how explicit intentions contribute to the formation of such a space of actions instead of maintaining a strict distinction between these explicit intentions and the space of actions. Only then are we allowed: "usher habitual actions, or at least a subset of them, into the space of reasons. That subset will consist of those habitual actions which cohere with the agent's world view" (Pollard 2005 81).[303] A couple of characteristics can be formulated that would hold for such a framework. The first two characteristics are related to the mechanisms and the representations involved, while the other two characteristics refer to structural properties of the comprehensive result of these.

A first characteristic refers to the rather complex nature of the mechanisms involved. Based upon our earlier observations, we are not only expecting to find the relevance of a hierarchical structure of control, but we also expect to see how a sculpted space of actions is being employed. This space of actions will likely contain action representations at several levels of specificity and will have been sculpted over an extended period of time. That is to say, even though there is an important role for the top-down flow of control, this does not imply that at all times an action must be determined through the actual involvement of the comprehensive cascade of intentions. For one, as we've noted several times earlier, the hierarchical structure of complex and dynamical systems is in fact heterarchical, allowing the development of direct connections of upper with bottom levels of the hierarchy, with the evasion of intermediate levels (Berntson and Cacioppo 2008). For another, with a sculpted space of actions at several levels of specificity in place, not at all times do actions

---

[302] Concurring with Aristotle, such habituation of moral action can be considered as the development of a 'second nature', leading to the automatization of its performance (McDowell 1994). However, what needs to be warded off in that case, is the critique that such an action does no longer have its origin in the 'space of reasons' but merely in the 'space of causes', making it rather comparable to a hard-wired reflex. McDowell argues that Sellars – who introduced the notion of the two spaces (Sellars 1997) – made a strict distinction between the two, which he aims to tear down (McDowell 1994).

[303] With that, it can be argued, acting for reasons can even hold for those actions that we perform automatically or habitually, without necessarily always requiring preliminary reflection (Pollard 2003).

require still to be determined by the full intentional cascade. Perhaps only triggered by an unexpected orchestral intro the expert singer can switch from performing Don Giovanni to Saint François instantaneously, without needing to revisit his earlier reflections upon the interpretation of these roles while still using the corresponding modes of singing and acting. Instead, action and motor representations pertaining to these roles are already established and the appropriate ones can be activated through more or less consciously established intentions.

A second characteristic refers to the multiple representations that are involved in the determination of action. As we have learnt from the previous Part, in which we discussed development and learning, a single task can be carried out with the use of different representations, some of which are redescribed versions of others. In the present Part, this would imply that indeed more than just a singular representation of a particular action can be involved in the complex process leading to an action. Consequently, a challenge is the handling of these multiple representations, avoiding the influence of potentially incongruent representations. Or it may be difficult to specifically modify an action when it is performed habitually, for the underlying representation is then activated as a whole. It may be difficult for our expert singer to give his Don Giovanni a more androgyn pose, because he has over the years practised a rather masculine voice and pose, making it difficult to single out his pose for adjustment in another direction.

A third characteristic of the framework refers to the large productivity of actions that humans display and the differentiated role of certain action representations therein.[304] Remember that we've been discussing earlier the fact that in complex and dynamical systems we can expect some components to become generatively entrenched (Wimsatt 1986), rendering these components a more prominent role in the system's performances than others. A similar observation was made in the previous Part concerning the kludges that are formed in cognitive mechanisms. As for the present context, we may expect to find that some actions are involved more generally than others in the configuration of novel action.[305] Is it indeed the case that some actions or component actions, that are well practised and mastered, have indeed

---

[304] Indeed, even though great apes demonstrate mastery of complex and hierarchically structured actions and thus a potential for productivity in their actions as well, it may be a lack of motivation and curiosity that keeps them from demonstrating such productivity (Byrne and Bates 2007).

[305] Approaching this from a rather different perspective, Ricoeur has demanded attention for this capability in terms of the 'configuration' of action and the associated process of transfiguration that occurs both in performing and interpreting an action (Ricoeur 1991c ; Ricoeur 1992). Interestingly, Ricoeur also argues in favor of a certain hierarchy in the narration and consideration of action, which in turn can contribute to the organization and planning of life.

a greater probability of turning up in new configurations, of which the representations are figuring apparently more prominently in the agent's sculpted space of actions? Perhaps we can indeed find at several levels of specificity such deeply entrenched action representations, ranging from a particular vocal timbre to a more general mode of expressivity in an expert singer's performance, for example.

Finally, the framework's fourth characteristic refers to a structural property of the complex process of determination of an action. Given the diversity of component mechanisms and representations involved, a major concern could be how to secure a minimum amount of coherence and consistency between actions.[306] Indeed, an important consequence of the hierarchy implied in the intentional cascade model is that it fosters the coherence and consistency between actions. The question that presents itself is whether an agent's sculpted space of actions does in fact enhance or endanger this property. For we may fear an expert singer can at times be relatively easily misled into an inappropriate performance if his attention is diverted to an irrelevant stage prop because of the unintentional activation of a part of his sculpted space of actions. Does this relative autonomy of an agent's sculpted space of actions thwart the intentional cascade's support his coherent and consistent acting, or is it perhaps difficult to decide about this?

Scrutinizing these characteristics, we will navigate between a philosophical account of intentional action and cognitive neuroscientific evidence about action selection processes, consisting of information about the representations involved and the neural implementations. While doing so, we will look for the leeway with regard to explicit specification of action that the philosophical account of intentional action to be discussed offers us, such that the empirical evidence regarding intentional action concurs with it.

---

[306] This feature is related to the phenomenon of generative entrenchment, discussed in Part II. Once a stable feature becomes integrated in a complex mechanism, it is better able of developing – generating – new capabilities that build upon this entrenched feature, contributing to the mechanism's coherent performance (Wimsatt 1986 ; Wimsatt 2007).

## 2  MOTOR INTENTIONS: THE FIRST STEP IN THE HIERARCHY, OR NOT?

Building up our account of intentional action by starting from motor intentions and gradually developing it into a more complex and dynamic account, may raise some questions. Why, indeed, aren't we adopting the order of the cascade of intentions and don't we start with the distal – future directed – intentions? Shouldn't we agree with Bratman, who denies such priority of motor intentions by contending: "not that there are no present-directed intentions, but that to understand what intentions are we should begin by concentrating on the future-directed case. This is the methodological priority of future-directed intention" (Bratman 1984 379)? If an opera singer is to perform a particular role at a particular time and place at all, this intentional action can only occur if his distal intentions have priority over his proximal and motor intentions.

The nature or role of intentions is manifold, as was mentioned above. For Bratman and other philosophers of action, however, intentions are foremost 'terminators of practical reasoning', when reasoning has culminated in the formulation of an intention to act sometime in the future – near or distant. Another important role intentions play is as 'prompters of practical reasoning' when decisions about means-end relations have to be made, pertaining to a current situation. Third, they also contribute to the coordination and organization of action (Pacherie 2006). All three roles depend upon their being future-directed, rather than being directed only at a present situation. To the extent that intentions are necessary for coordination and organization of an agent's many intentions and actions, their distal versions are even more crucial. Proximal and motor intentions do not explicitly integrate information about moments and situations that transcend a current situation, diminishing their role in such coordination. For that reason mainly they're not given the methodological priority that is lent to distal intentions.

However, this apparent simplicity of especially motor intentions is precisely our reason for taking these as a starter. Given the importance in our account of learning and development and our interest in the increase of complexity of the mechanisms that can be used to explain action, starting with distal intentions would be odd. Indeed, when we follow the order of intentions as they emerge during ontogenetic or phylogenetic development, we can better explain why we can expect this development of a hierarchical structure and the differential generative entrenchment of intentions to occur. For with the increase of capabilities for actions, there is also an increasing demand for saving the cognitive resources required for their performance and for avoiding the performance of incoherent or even contraproductive actions. By starting

at the lower end of the intentional cascade, therefore, we aim to show that its increasing complexity and dynamics – including inter-level interactios - is not surprising, as soon as some form of learning and development is taking place.

To be sure, this set-up of an argument about development or increasing complexity is not new – not even in philosophy. Aristotle, for example, presented in his De Anima an analysis of different kinds of 'souls' or their functions, which are related such that the analysis of the most complex – human – soul can build upon the analyses of the more simple vegetative, sensory, and locomotive souls, as their capabilities are integrated in the former one.[307] Although driven by very different ambitions, Hegel followed Aristotle when providing a comparable trajectory by starting his Phenomenology of Spirit with an analysis of perceptual certainty (sinnliche Gewissheit) that included no explicitly articulated contents (Hegel 1988). Belonging to another tradition, we can find arguments for such a set-up that to some extent hold for the present one, as well.

In his discussion of method in philosophical psychology, Grice defends what he calls: 'creature construction'. Employed as a heuristic, he engages in this creature construction by describing stages of a creature that is capable of having increasingly complex mental states. Applying psychological concepts in this way, one allegedly can: "compare what one thus generates with the psychological concepts we apply to suitably related actual creatures, and when inadequacies appear, to go back to the drawing-board to extend or emend the construction" (Grice 1974 37). Inspired by this heuristic, Bratman follows suit by constructing even eight different creatures with increasingly hierarchical planning structures, contending that Creature 1 is moved merely by the strongest of his first order desires and can therefore hardly be called an agent, while Creature 8's coherent and consistent planning is facilitated by hierarchical and feed back structures (Bratman 2006b, ch. 3). Although Creature 8 does not lack Creature 1's strong desires, the former has meanwhile reached such complexity that these strong desires can be put to work while being coordinated with other constraints. Something alike will be described in the present context.

Our approach does not involve the construction of hypothetical creatures, but is in nature more akin to the dynamical or developmental approaches mentioned earlier. In terms of our returning example, we will not start with the analysis of the opera's singer comprehensive interpretation of a role but devote this first analysis to the intentional control of ongoing motor actions. The analysis will especially focus on the increasing complexity and automaticity with which such actions can be performed. Based upon

---

[307] Aristotle's psychology exemplifies how a materialist – or mechanical – explanatory strategie can be combined with a teleological one, by underlining how simple bodily functions are reorganized and take up new roles when integrated with more complex cognitive functions (cf. (van der Eijk 1997).

these properties, an opera singer can increasingly expand his repertory of complex performances and adjust these to a current stage setting at wish. As a result of dynamic properties of motor intentions, therefore, the agent is capable of further sculpting the space of actions such that proximal and distal intentions are facilitated, if they somehow build upon the emerging properties of this sculpted space.

In the next two sections we will look more closely at these motor intentions, observing whether the four characteristics listed in the previous section - section III.1.3. - figure in philosophical and cognitive scientific accounts respectively as we would expect them to, based upon our previous analyses. For our expectation is that an agent should be able to produce ever more complex behavior as a result of some sort of kludge formation which also affects the action representations involved.

## 2.1  A philosophical analysis of motor intentions and guidance

When for our first analysis we aim to separate motor intentions from the kinds of choices or decisions that are the contents of proximal and distal intentions, are we not left with mere physical movements? Are goals and criteria for satisfaction of an action not set before an action unfolds in a cognitive process that is separate from that action itself? Such considerations have partly motivated the development of causal theories of action. Davidson, for example, did compare the action to a causal event with the difference being the fact that in our description of an action we refer to an agent's reason for action as its case, leaving the nature of the subsequent action relatively untouched by its preceding intention (Davidson 1963).[308]

Such theories have drawn several lines of critique which we must leave aside, since this is not the place for a detailed discussion of this debate.[309] Based upon philosophical

---

[308] Even though Davidson compares his approach to Aristotle's comparison of actions and events in the Physics, there is reason to reject that comparison. Aristotle is more liberal in admitting different kinds of determining factors and also clearly acknowledges that such a factor must not always precede its effect, making Aristotle not vulnerable to a criticism that has been raised against the causal theory of action. (cf. Sorabji 1980 ch. 2).

[309] Although our focus is on Frankfurt's and Bratman's analysis, other authors have contributed to the debate. Searle, for example, has introduced the helpful notions of prior intention and intention-in-action in order to account for the distinction between an intention that precedes and causes an action, and an intention in action that has been caused by this prior intention and is itself more responsible for determinating appropriate bodily movements (Searle 1980). Nonetheless, Pacherie writes critically of such 'dual-intention theories' that "they tend to assume that the role of the first of these two intentions is over once the second is in place" and "that action guidance and monitoring are the sole respnsibility of the second intentions" (Pacherie 2008 181). This is not correct. Searle, for instance, has agreed to have made that mistake and accordingly later added to the causal relation of a prior intention to action also a constitutive relation (see Searle 2001 51, n. 5). Bratman's future-directed intentions are explicitly meant to retain their coordinating role once they've become present-directed intentions (Bratman 1992b). Finally, Mele states explicitly that he'll argue that "the moving role of proximal intention extends beyond triggering to the causal sustaining of the functioning of actual mechanisms" (Mele 1992 173).

and empirical considerations our aim is to convince the reader that such a separation between an agent's intentions and his corresponding action is implausible. If we can observe experienced agents engage in complex and flexible motor actions that are carefully planned and exercised, than it seems plausible that the momentary motor movements can be deeply shaped and influenced by intentions – even if they operate at different temporal scales and are determined in terms different from the 'metrics' of those movements. If this is the case, however, then we need to show that motor movements are intentional not only in the sense of unfolding under the ongoing influence of intentions. On top of that, we must ask if and how these 'motor intentions' can be shaped under the influence of those other - proximal and distal – intentions. In short, in contrast to a causal theory of action, our aim is to show that the apparent simplicity of an unfolding action hides a complex and dynamic intentional nature.

Responding to the causal theory of action of Davidson and others, Frankfurt has emphasized how both the phenomenology of action and its analysis betray that an action remains determined by agent's intentions until the very end. His approach was taken up by Pacherie when developing her notion of motor intentions, as she agrees with Frankfurt's insights, who: "argues that what distinguishes an action from a mere bodily movement is the fact that the person is in some particular relation to the movements of his body during the time in which he's performing them and that this relation is one of guidance" (Pacherie 2008 190). She then goes on to interpret Frankfurt's insights in terms of motor control at the lowest level of the intentional cascade.[310] Let us look whether their combined account does include the necessary complexity of motor intentions, being not only responsible for ongoing control but also being capable of modifiability via processes that are under the control of those other intentions – proximal and distal intentions. Moreover, let us notice whether the account includes also a form of differential generative entrenchment of the motor intentions, allowing some a greater role than others in complex motor actions.

Taking up the alleged separation between an action and its preceding intentional cause, Frankfurt clearly defines what is at stake: "[t]he problem of action is to explicate the contrast between what an agent does and what merely happens to him, or between the bodily movements that he makes and those that occur without his making them" (Frankfurt 1988 69). Admittedly, most specific bodily movements are not intended in detail by an agent, yet it is problematic to maintain that they are only happening to him

---

[310] In fact, Pacherie correctly notes that philosophers have not always clearly distinguished between two forms of guidance, which she aims to correct: she distinguishes between higher- and lower-level guidance and monitoring, corresponding to proximal and motor intentions and having correspondingly different properties (Pacherie 2006). Frankfurt does also not distinguish between different levels of guidance but we'll focus here on those properties of guidance that are more strongly related to motor intentions.

and not guided by him.[311] On the contrary, we usually assume that if an agent is acting, he: "must be in some particular relation to the movements of his body *during* the period of time in which he is presumed to be performing an action" (Frankfurt 1988 70, italics in original). This relation is not identical with the kinds of intentionality that are located at higher levels of the intentional cascade while it seems to be lacking completely in cases when the agent's pupil dilates in response to incoming light or when he suffers from a muscular spasm. Instead, for the guidance that is involved in action we can better consider cases when ongoing movements that appear to be made automatically still demonstrate coherently many complex and meaningful patterns, as when a musician is guiding his body parts (Frankfurt 1988 72).

Such ongoing guidance of action is not usually dependent upon an agent's conscious decision making, which forces us to ascribe guidance to a drug addict's behavior both when he is taking his drugs because of his addiction or upon his free choice (Frankfurt 1988 76). Indeed, guidance is often necessitated by mere changes in movement or in the environment that threaten to impede successful completion of the behavior. Frankfurt suggests that mechanisms responsible for guidance may linger in the background when an action is performed, being prepared to intervene if necessary. He compares this with a driver who only then intervenes when the speed or direction of his vehicle no longer satisfy goals or criteria that have been set previously. In such a case, the guidance mechanisms may remain passive when: "no negative feedback of the sort that would trigger their compensatory activity may occur" (Frankfurt 1988 75). Let us first look closer at the relation between an agent and his guided action, realizing that the nature the guidance relation is partly dependent upon the information or contents involved. Indeed, purposive guidance can only occur after some contents of behavior have become particularly relevant to the agent.

Take again the case of the musician, whose movements do not simply 'happen' to him as they are meaningful and coherent patterns that result from extended periods of intentional practice (Frankfurt 1988 72). Due to such practice, he is so familiar with the patterns that should be performed and the sounds that result from these, that negative feedback may occur when the comparison of these patterns with the actually performed body movements or musical sounds yields incompatible results. Such situations arise, for example, when the musician strikes a dissonance or when his tempo is not in sync with the orchestra. Often in such cases, as experienced musicians

---

[311] It seems to me that Aristotle has already put an elaborate conceptual apparatus in place to account for several phenomena that require nuanced explanations. For example, Aristotle could make a difference between cases in which the cause of a movement is 'in the agent' but is still not 'upto him' implying that he has no cognitive control over the movement, as is the case with certain pathologies or intoxication. Cf. (Sorabji 1980).

know, it is often hard to resist repeating the failed note or adjust automatically one's tempo in order to correct the mistake. So guidance is activated in cases when a complex and purposive behavioral pattern is either inappropriately or incorrectly performed or not in harmony with relevant environmental conditions and then issues in interrupting and perhaps repeating the attempted behavior.[312] Similar to the high way being a car driver's environment, we may perhaps consider the music score – remembered, or not – as an environmental condition for a musician that presents conditions for the activation of guidance mechanisms in order to adjust his dynamics or tempo. However, not all behavioral responses to environmental stimuli should be considered as involving guidance.

Naturally, it is not to be denied that pupil dilation is a purposive movement of some body part. However, if we refer to guidance at all in such a case, this guidance: "is attributable only to the operation of some mechanism with which he cannot be identified" (Frankfurt 1988 73). Identifiability with the guidance mechanism is apparently an important characteristic for guidance to be not just purposive but genuinely intentional. Perhaps, so we may ask, this identifiability is associated with the long-term processes that allow the modification of motor intentions that we expect to find in the account? At least we have learnt that according to Frankfurt, first, many behaviors are accompanied by some distinct mechanism that is responsible for an agent's guidance of it. Second, such a guidance mechanism can be more or less connected to an agent's identity. The examples of the driver, the drug addict and the musician suggest that we can observe guidance mechanisms with which an agent is identifiable at work in their actions and not in pupil dilation, the difference being that the latter is a motor reflex while the former examples refer to habitual and long-term intentional actions. Those actions which gradually develop under the influence of proximal and distal intentions and are characterized by the richness of their contents, compared to pupil dilation movements.

The richness of intentional action is discernible from the goals, relevant environmental information and criteria that are relevant for its guidance. This same richness is related to the identifiability of the agent with such action. Several years after his earlier account of action, Frankfurt further embeds his notion of guided behavior in the context of the agent's overall constitution.[313] For then he argues that

---

[312] Although approaching Pacherie's model of agency from a rather phenomenological perspective, Gallagher concurs that for adjusting our motor intentions we do not always need to rely on higher level intentions (Gallagher 2012).

[313] Nonetheless, it has been argued that Frankfurt's overall position can be characterized by his emphasis on the phenomenon of guidance. Indeed, it is this notion that is crucial in his opposition to the causal theory of action, according to (Di Nucci 2011). Di Nucci stresses that guidance is relevant for all forms of skilled, habitual and routine activities, which concurs with our view here.

the performance of an action depends upon the agent's wholeheartedness with respect to that action, in which case his intention to act has co-determined: "his cognitive, affective, attitudinal, and behavioral processes" (Frankfurt 1999 103).[314] Such a wide range of processes that involve various kinds of information, we may note, are obviously not involved in pupil dilation nor in the purely physical event that an action was deemed to be according to the causal theory of action.

Given both the informational richness and the multiplicity of the processes involved in the development of this kind of guidance, it seems plausible that it needs more time and attention to arise than motor reflexes do. Indeed, such guidance is usually an effect of an agent having gained experience with or deliberately learnt a particular meaningful pattern of movements.[315] Such guidance is not involved in the movements – complicated as they are - that an epileptic patient makes, for it is: "unlikely that a person would have created such an incoherently complicated pattern if he had been guiding his body through its movements" (Frankfurt 1988 72). In contrast to such a patient, we can observe in guided behavior that is a result of practise and experience a relation between someone's – or an animal's – persistence and the consistency of his performance that results from it, as guidance does entail: "a certain consistency or steadiness of behavior; and this presupposes some degree of persistence" (Frankfurt 1988 84). Again, the musician's consistent performance of a complex piece of music after having practiced the score for many times is exemplary.

In sum, according to Frankfurt's analysis guidance is at stake whenever an agent is performing a complicated and meaningful pattern of movements, during which his enduring intentions play a certain role. Relying on separate mechanisms– being different from the mechanisms that produce ongoing movements – and employing relevant information about aim, meaning and context of the action, these mechanisms only intervene when it is necessary to adjust the action, while remaining passively in the background for the rest of the time. The analysis does not offer detailed insight about the information that is used for guidance but only contends that guidance

---

[314] In his analysis of the problem of action, a similar view was already announced when Frankfurt wrote that: "[t]he facts that we are rational and self-conscious substantially affect the character of our behavior and the ways in which our actions are integrated into our lives" (Frankfurt 1988 77). Behavior being guided is an important aspect of its character in this context.

[315] Guidance is different for instrumental actions directed at a goal-state and for actions that are performed for their own sake. For example, Aristotle explicitly ascribes to music such an intrinsic value and distinguishes music making from actions that have an external goal, even though music education can have beneficial effects on the pupil's character (Politics, VIII, 5-7). In the next chapter we will take note of the fact that in cognitive scientific investigations of intentional action, such action is usually defined in terms of its goal-directedness, often with the involvement of an associated external object. Nonetheless, musical processing is being studied precisely because it is possible to investigate the role of meaningful patterns in the absence of a specific (end) goal of the behavior, comparable to language processing (Levitin and Menon 2003).

stems from consistent practice of patterns of movements during an extended period of time. In any case, guidance appears for its interventions to not necessarily require information about the goal of an action or about higher levels of intention regarding the action.

Given our previous analyses of dynamical modifications of underlying mechanisms, of kludge formation and the role of modifiable – redescribable – representations in these, let us note a couple of issues that merit further analysis. To begin with, it seems that the analysis distinguishes between guidance of intentionally initiated behavior and bodily movements that merely happen to him without including a third option: highly automatized actions that are guided but the initiation of which may at times not be under the agent's control, as when a musical attack is triggered by the environmental stimulus provided by the accompaniment.[316] We can only suggest at this point that these options probably differ more in degree than in kind, sharing properties and underlying processes with each other. Deciding about that suggestion would require further clarification of how the patterns involved in guidance are stored and retrieved, as would be an answer to the question whether the representations involved are somehow redescribed during learning – an algorithmic theory in Marr's sense (Marr 1982) remains to be formulated. Finally and related to this, it merits further discussion how experience with practiced patterns of movements may enhance the agent's consistency in his performances, as has been suggested. Is practice affecting the representation of these complex patterns and is it more specifically thanks to a hierarchical structure that an agent's actions increase in coherence and consistency – as we will learn from other analyses?[317] Although we've suggested that some of these issues are relevant in the context of Frankfurt's analyses, turning to a more empirical perspective on motor intentions will perhaps provide more detailed insights – concurring hopefully with some results of our earlier analyses.

---

[316] In his analysis of 'identification and externality', Frankfurt also distinguishes between passions that are internal or external to a person. There, however, he points out that a person may come to recognize a certain passion as his own, even though he regrets having that passion (cf. Frankfurt 1988 64). Could we assume that similar to this situation, an agent may observe that some of his automatized actions are no longer under his comprehensive control even though they are clearly attributable to him? Frankfurt seems to deny this, when he elsewhere protests against Aristotle's view on a person's responsibility for his own character and argues that Aristotle overemphasizes causal responsibility for his character instead of considering rational identification with his character as (cf. Frankfurt 1988 171). Frankfurt appears to overlook, however, the fundamental place of action in Aristotle's ethics in general, including his view on character. Indeed, this view on character implies that one is in a position to influence it – perhaps after having identified with some rather than other characteristics – by acting repeatedly in line with those preferred characteristics and by trying to increase control over others through practice as well. See also (Audi 1991) on this issue.

## 2.2  Motor intentions and chunks: evidence about developing complexity at the bottom of the hierarchy

Looking for motor intentions in the analysis of action according to the philosophical hierarchical account that provided a basis of Pacherie's intentional cascade, we learnt that these were particularly implied in the phenomenon of behavioral guidance. This parallel notwithstanding, we looked in vain for some aspects of motor intentions that we did expect to find on the basis of our previous analyses of modifications of complex and dynamic mechanisms, of kludge formation and of the redescription of repres s are involved in  development and learning, as is the case when babbling babies grow into expert singers. Let us consider whether empirical research does provide evidence for these aspects of motor intentions. In order to do so, let us first consider what empirical research Pacherie's model refers to.

Pacherie explicitly notes that motor intentions differ from the other two kinds - proximal and distal – of intentions. The contents of motor intentions are not propositional, for example, and they conform to some features of modularity, as they are to some extent informationally encapsulated and cognitively impenetrable (Pacherie 2008 187 ff.).[318] Notwithstanding these differences in information processing, motor intentions are an integral component of the intentional cascade that is further elaborated as a complex model of action control. Indeed, it is suggested: "that the information-processing model of action control in terms of internal models be explicitly combined with the threefold distinction among levels of intentions […], thus yielding a richer theoretical framework for thinking about action" (Pacherie 2008 193). Allegedly, action control relies on the use of models of action that are employed by the processes involved. What representations and what kinds of information processing are implied here?

---

[317] Sharing the interest in a hierarchical account of agency with Frankfurt, Bratman has barely touched upon the topic of motor intentions in Pacherie's or guidance in Frankfurt's sense. He seems content to note that the specific motor movements that are 'necessary constitutive means' of an intended action belong to the motivational potential of that intention, while they need not be articulated by that intention. For example, the intention of shooting a jump shot may well imply as a means to that end that the agent must stop on his left foot, even though the latter motor intention – in our terms – is not included in the intention: "motivational potential can be extended by means-end beliefs, even when what is intended is not thereby extended" (Bratman 1984 401). Indeed, other beliefs and desires may play a role in this motor movement with his left foot, for example the agent's necessity to spare his hurt right foot. Clearly, this does not yet present a role for independent guidance mechanisms that somehow adjust those constitutive motor movements by employing a representation of meaningful and coherent patterns of movements that have been gathered by the agent.

[318] The non-conceptual nature of the representations involved in motor intentions does not preclude their having properties usually ascribed to conceptual representations. For example, non-conceptual representations have conditions for their satisfaction; there are misrepresentations possible; non-conceptual representations can be composed and have a hierarchical structure, too. See (Bermúdez and Cahen 2012 ; Bermúdez 1995) for a discussion of non-conceptual content in cognitive functions.

From the available accounts of internal models of action involved in intending, monitoring, guiding and adjusting motor actions, Pacherie's framework leans in particular upon the account of motor representations presented by Jeannerod and others (Pacherie 2000 ; Pacherie 2008). Drawing together research on motor action, motor imagery and observation of motor action, Jeannerod had earlier argued that the cognitive processes involved in these tasks make use of largely overlapping 'motor representations' which are not limited to representations of motor contractions.[319] On the contrary, these representations merit independent attention when studying overt action because these: "representations are likely to be endowed with properties (partly built on experience from previous actions) which may not be apparent in their eventual motor counterpart. They seem to be structured with different levels of organization; they use cognitive rules for establishing the serial order of action parts, for assembling programs, etc." (Jeannerod 1994 201). It is the complexity of representations and their associated modifiability that will be found especially important for motor intentions and deeply involved in the process of kludge formation that is observable with regard to motor intentions, too.[320]

Just like Frankfurt did notice that experience and practice have an impact on guidance, so did he note that guidance does not require language. Jeannerod does ascribe to that position, as well. Although he does not assign to motor intentions all characteristics of modularity as presented by Fodor nor its characteristic of informational encapsulation alone (cf. (Fodor 1983)), Jeannerod does make a distinction between semantic and pragmatic modes of knowledge regarding actions and motor movements. In his account, these modes correspond largely with the two different streams of visual processing, which he distinguishes as semantic and pragmatic. However, he does acknowledge that there are still interactions between those streams (Jeannerod 1997 ; Jeannerod 2006).[321] Notwithstanding these interactions, this pragmatic mode of knowledge is in itself non-conceptual and is accessible for explicitly articulated and

---

[319] Theoretically, sensory representations could have occupied a similar central role as the motor representations in this framework. However, evidence from different lines of research shows that the representations involved in a wide range of tasks related to motor behavior, like its perception, imagination, performance and verbal expression, are subject to modifications as a result of experience with their performance and not as a result of perceptual experience with them (de Vignemont and Haggard 2008). Similarly, the shared representations of action that play a role in the activities of mirror neuron systems which are activated when subjects try to understand, imitate or perceive an action appear to be primarily sensitive to motor experiences a subject has gained. Indeed, representation of perceptual information about an action only would provide little for understanding that action, it is argued (Rizzolatti and Sinigaglia 2008).

[320] Jeannerod and others confirm the modifiability and complexity of both perceptual and motor schemas when they observe: "that new schemas often arise as modulators of existing schemas rather than as new systems with independent functional roles" (Jeannerod, Arbib et al. 1995 361).

[321] About the number and interpretation of visual streams, discussion is still going on – see footnote 50.

conscious modification only indirectly, for instance via motor imagery (Jeannerod 1997).[322] Apart from this limitation, motor representations are complex, flexible and dynamic and subject to indirect adjustments via other cognitive processes and under the influence of experience. Kludge formation, therefore, can possibly involve these motor intentions, too. We will come back to that later, but must first mention an important role that these motor representations play in Pacherie's framework.

That framework integrates the motor representations with other components that together are responsible for the specification of motor actions and the multiple forms of control involved in these. For, corresponding to the framework's different levels of intentions it includes also different levels of motor representations. Moreover, during the processes involved, such representations are used both for the representation of the desired, predicted and actual states of the relevant behavior and include representations of relevant properties of the objects and the situation for action. In fact, it is the complex process of comparison between those states at the three different levels of specification that facilitates control of the action and also contributes to the agent's sense of control. For example, when a motor intention that specifies the movements to grasp a certain object leads to movements that eventually miss the object, comparison between the relevant representations may result in a sense of lacking motor control – and usually in a correction of movement (Pacherie 2008).[323]

Given this elaborate structure of a hierarchy that includes the intentional cascade, multiple representations and several relations between those, the question presents itself whether this structure also allows for the four characteristics of a comprehensive framework for the explanation of action in which a sculpted space of actions plays a role, as formulated in section III.1.3? Remember, these characteristics were formulated in line with our interest in the explanation of action and its potential modifications under the influence of development, learning and expertise. They referred to potential

---

[322] Jeannerod also elsewhere refers to the fact that both experiments and patient studies confirm that automatic and implicit processing of motor action usually functions in relative isolation from conscious and explicit – verbal – processing of the representations involved in such motor actions. The former is also often spared in patients who have difficulty with conscious and explicit processing of information related to motor action, with the reverse occurring more seldomly (Jeannerod 1997 ; Jeannerod 1999). Other evidence of the distinction of the two modes of processing comes from experiments like those with subjects who automatically adjust their grasping movements to a changing objects, before being aware of any changes (Jeannerod 1997).

[323] The comparator aspects of the framework build upon Wolpert's a.o. work on the contribution of forward and inverse models on neural processes, responsible for planning, control and learning (Wolpert, Ghahramani et al. 1995 ; Wolpert and Ghahramani 2000). Jeannerod has similarly used Wolpert's work to develop his theory of motor simulation, which explains how similar motor representations are used in different cognitive functions – like action, imagination, observation, verbal expression – and also contribute to many different cognitive phenomena, including pathologies or other surprising phenomena (Jeannerod 2001 ; Jeannerod 2006).

changes in the underlying mechanism that leads to an action and the mechanism's employment of different representations, to the differential generative entrenchment of some representations and the involvement of these in new actions, and finally to the demands for coherence and consistency in action which should be fulfilled partly by the structure's hierarchy. Did the philosophical analysis of motor intentions and guidance present limited confirmation of these characteristics, let us now look for further confirmation in this section on empirical research on motor intentions and focus particularly on the role of expertise in that context. Doing so, we will devote separate sections to the connection between expertise and motor intentions, to modifying mechanisms, to the differential generative entrenchment of motor intentions and finally to the coherence and consistency between actions.

### 2.2.1 Motor intentions, representations and the role of expertise

Even though motor intentions are ascribed a relative autonomy from the other intentions, they should enable an agent to act appropriately in a particular situation.[324] Motor intentions as contained in the framework enable an agent to do this by establishing associations between perceived environmental conditions and an appropriate motor response to these. Motor intentions are said to be involved in the 'pragmatical organization' of a perceived situation by an agent, who can demonstrate how: "affordances of an object or situation are automatically detected even in the absence of any intention to act." As mentioned, it indeed is not just a matter of the detection of such an affordance, though, as this is directly connected to potential actions, since: "[t]hese affordances automatically prepotentiate corresponding motor programs" (Pacherie 2008: 186).[325] So even though the term suggests otherwise, motor intentions involve not just representations of potential motor actions but also of relevant environmental properties that are associated with these actions.[326] A related issue that will be discussed in the next section on potential mechanism modifications

---

[324] In utilization behavior, agents act inappropriately as upon the perception of a specific object an agent is unable to inhibit an action afforded by it, even if the result of the action is undesirable – as is the case when a second set of glasses is put on (Lhermitte 1983). This pathology is associated with lesions in the frontal lobes (Lhermitte, Pillon et al. 1986). It underlines the complexity of expertise, which is more than just stimulus driven, reflex-like action as that would oftentimes be inappropriate or ineffective.

[325] Jeannerod explicitly remarks that affordances are among other things those properties of objects that afford specific motor patterns, without necessarily offering cues for a perceptual category (Jeannerod 1994). The fact that many event codes or features – like those of spatial orientation - involved in the representation of perceived information and in motor control are shared implies that not for all affordance features a comprehensive cognitive transformation is necessary in order to get from scene perception to motor action (Hommel, Musseler et al. 2001). Particularly neurons in cortical association areas are involved in coding for features pertaining to both perceptual, and cognitive and motor functions related to a single task (Cisek and Kalaska 2010).

associated with motor intentions and expertise is how motor intentions are learnt and stored at all. For now, we can expect that the automatic detection of action options in a specific situation and the automatic prepotentiation of associated motor programs is subject to change through expertise.

Given that any situation affords many options for action and that there are usually multiple motor movements available in order to realize a single option, the potential space of actions is large and expertise may be expected to play a role in sculpting that space. A domain of action that allows researchers investigation of this expertise-dependency, as it allows the calculation of the number of options or the outcomes of different options, is chess. Indeed, the best researched domain of expertise and associated cognitive functions is perhaps the domain of chess. Much cognitive psychological research has been conducted on the acquisition of skill or expertise in chess since De Groot's seminal studies (de Groot 1946).[327] Even though different explanations have been offered, researchers agree that expertise in chess is associated with a broad – and connected - variety of improved capabilities like enhanced perception, recognition, and storage of patterns or board positions, improved envisioning of potential responses and finally better results in acting (Chase and Simon 1973 ; Dreyfus 2004 ; Ericsson and Roring 2007).

Nonetheless, it may be contended that since chess involves rather limited motor activities, its generalizability is perhaps limited. However, the generalizability of chess is supported by the fact that even though cognitive and perceptual-motor skills are usually investigated separately, there are several arguments not to treat them strictly separate. For there is considerable overlap in their acquisition and in brain activations activated with them, and it appears that in all cases a distinction can be made between explicit and implicit knowledge involved in these skills (Rosenbaum, Carlson et al. 2001). Moreover, studies of expertise in domains in which physical action is required

---

[326] A related but still different issue is whether intentional motor action does always imply a specified goal. Humans appear in general to be 'obsessed with goals' as some investigators concluding and associating it with the prevalence of a teleological stance in action observation (Csibra and Gergely 2007). Indeed, scientists who investigate intentional action appear to share that obsession, as most actions studied include them having clear goals. However, action goals can be various and include an agent's end state, both postural and motivational – for example one's goal for running may be to lose weight, which refers to a desirable end state. Nonetheless, defining such distinct actions goal may also be a matter of different levels of specification, since a runner must in any case specify the direction and goal for running in order to implicitly determine the necessary motor intentions. Simplified to the extreme, observers can also imitate meaningless actions by encoding only the spatiotemperal layout of the necessary movements (Decety, Grèzes et al. 1997). It is a matter of definition, if one wants to call the latter goal-directed behavior, too.

[327] Gobet lists nine reasons why chess is such a profitable domain for the study of expertise, among which are its ecological validity, its offering the ELO ratings of players as a quantification of expertise, its use in artificial intelligence experiments, and of course the complexity of the tasks involved (Gobet 1998).

show improvements that are comparable to those that have been found in studies of chess expertise, as demonstrate studies in nursing (Benner 2004), in fire fighting (Klein, Calderwood et al. 1986), or in sports (Janelle and Hillman 2003). Moreover, it is also commonly acknowledged that these improved capabilities do in real situations not rely on conceptual or declarative knowledge, but rely instead on a different type of knowledge, often referred to in the literature as 'intuition'. Supported by these arguments, let us therefore take a closer look to how chess experts deal with the contents of their domain of expertise and how this is different from novices and let us do so by focusing upon an explanatory approach that aims to combine insights of several other approaches: template theory (Gobet and Chassy 2009 ; Gobet and Simon 1996).

Building upon De Groot's findings, research of perceptual and memory skills and their interaction in chess experts showed that they are capable of processing much more information than novices do and recognize relevant information much easier, doing this also faster and more reliable. Experiments in which players have been exposed to complex positions for a limited amount of time, after which the players had to reproduce from memory these positions, for example, demonstrated significant differences in accuracy and completeness between experts and novices. As experts face cognitive and memory limitations identical with those that novices face, the differences have been explained by hypothesizing that experts automatically enlarge the amount of information that can be processed at a time by chunking and thus condensing it (Chase and Simon 1973). Indeed, it appears that where novices have difficulty in recognizing and processing meaningful patterns, experts easily group stimuli in meaningful groups also called chunks, a chunk being defined as: "a collection of elements having strong associations with one another, but weak associations with elements within other chunks" (Gobet, Lane et al. 2001 236).[328] For example, a series of letters is difficult to memorize unless one can chunk that information with the use of memorized words: the letter series "andramoiennepe" can be chunked as an ancient Greek sentence of three words.[329] Indeed, it is the exhortation from the start of the Odyssey, which is followed by some 12.000 verses that have many times been

---

[328] This definition of a chunk resonates somewhat with the definition of a module, of which the internal relations or interactions between its elements are strong, while relations to other system components is weaker (Mitchell 2006).

[329] Parry was one of the first to draw attention to this accomplishment, pointing out that such oral traditions still existed in modernity. Moreover, he also referred to the role of returning formulas, descriptions and the like – chunks of information, as we may now call them - as 'singer's rests' in these challenging songs (Parry 1930). Bloch argues for the importance of recognizing that much cultural expertise is stored in 'chunked and non-sentential knowledge' and therefore challenging to anthropologists who aim to render it in linguistic form (Bloch 1991).

recited by heart by illiterate persons – a task that equally relies on their ability to chunk information at several levels of grain: as words, as sentences, as scene descriptions, and so on.[330]

Having improved capabilities of perceiving, recognizing and memorizing relevant situational conditions, experts have in addition gathered many associated, appropriate behavioral responses to each of these chunks of information.[331] An early account of such associations was in terms of the 'production' of an action in response to the perception and recognition of a specific pattern which happens to satisfy certain conditions (Newell and Simon 1972).[332] To begin with, an expert must discover at an early stage relevant situational conditions, as chess experts do who are quicker to recognize the relevant strategic patterns present in a certain board position (Ferrari, Didierjean et al. 2008).[333] Subsequently, experts generate or activate proper options for action upon the recognition of such patterns without requiring time-consuming, intermediate deliberation.[334] Indeed, this process of the generation of an appropriate action option is already initiated during the perception of the scene. For in sports as diverse as chess and basketball, expert perception is shown to be partly anticipatory in character as experts tend to focus their attention on positions that are particularly

---

[330] It is now estimated that expert chess players can memorize ca. 300.000 configurations of chess positions, ranging from complete boards to smaller configurations (Gobet and Simon 2000).

[331] Reviewing the literature on expertise, the authors list among the benefits of having expertise: having better perceptual and recognition skills, a larger set of routines, better ability of mental simulation and detection of problems and opportunities, more declarative knowledge about the skill and metacognition about his own capabilities and limitations (Phillips, Klein et al. 2008).

[332] In his ACT theory, Anderson posits that the interaction of production rules and chunks of declarative knowledge can together account for cognition. Perhaps, the declarative stage distinguished as the first stage of skill acquisition by Anderson applies to cognitive skills more than to motor skills (Anderson 1982).

[333] As reviewed in (Didierjean and Marmèche 2005), during perception unexpected – a drum in the kitchen - or un-anticipated – a change of direction of a moving object - scenes attract observers' attention, testifying to the fact that perception is not the passive process it is often made out ot be. In line with research of such effects of familiarity, a review of expert effects in memory recall in many different domains of expertise – ranging from chess and computer programming via medical expertise to sports – is concluded with a theory in which the authors propose that experts become attuned to the specific constraints that are relevant to the domain of expertise. They explicitly note the affinity of their theory to the ecological theory of perception with its emphasis on situational affordances (Vicente and Wang 1998).

[334] This is not to deny that deliberate practice plays an important role in the process of skill acquisition for experts, too. However, over time important physiological and functional adaptations have taken place that are associated with modified mechanisms if compared with novices (Ericsson and Roring 2007).

[335] Such an interaction between expertise, recognition of domain relevant situation features and expert's attention to these has been shown not just with memory recall but also with eye tracking experiments, for example in pilots (Bellenkes, Wickens et al. 1997). In a review of behavioral and neural correlates with decision making in response to external cues or rewards and to internal preferences respectively, the authors draw two conclusions. First, the two neural networks that are associated with externally and internally guided decision-making are not completely distinct but activations differ rather gradually from each other. Second, there is a large overlap between the internally guided decision-making network and the so-called resting state or default mode network, suggesting that internally guided decision making relies more on agents' internal preferences or criteria that are established over a longer period of time (Nakao, Ohira et al. 2012). Expertise, one may surmise, has lasting effects on such preference or criteria, too.

relevant for future moves (Didierjean and Marmèche 2005).[335]Assisted by that focus, the generation of appropriate action options is a less demanding task than if no such pre-selection of relevant situational features had occurred. Nonetheless, even with these constraints in place there are still several action options being generated in parallel, as several behavioral and computational studies suggest. From these, eventually a single option is being selected for execution, the selection of which is usually done implicitly (Cisek 2007). Expert accounts of their performance appear to concur with this account. For example, fire fighters who have been operating in real emergency situations and under time constraints, report afterwards that they did not compare in evaluative terms several options for action but instead experienced that one option simply stood out or presented itself as the most viable one in the current situation (Klein, Calderwood et al. 1986).[336]

What we've learnt from the above and will be confirmed more below, is that an expert's sculpted space of actions influences several components of the mechanisms involved in his performance. Based upon his many previous experiences, he has an enhanced capacity of perceiving, recognizing and focusing upon patterns of relevant situational features.[337] Moreover, the representations involved in these processes have become strongly associated with options for action, which in turn also influence the expert's focus of attention and subsequent perception of the situation. In that sense, the expert's sculpted space of action is affecting all processes involved. A matter of concern is to what extent this sculpting process may not only be enhancing the expert's performance but may also be limiting it, as it is perhaps inflexibly constrained by the specifics of the representations involved: does expertise only contribute to a facilitated repetition of the expert's previous experiences? Several lines of evidence and theories suggest otherwise, as we'll argue below. In this section we will focus on how the representations that are involved in expertise are structured such that they offer the kind of flexibility and adaptivity that we expert from an expert.

---

[336] Based upon an analysis of Conan Doyle's texts, it is stipulated that Sherlock Holmes relies upon similar mechanisms facilitating his expertise as described here. The authors point out that Holmes' case suggests some issues to be taken seriously in the study of expertise, among which the role of emotion in expertise (Didierjean and Gobet 2008).

[337] Research with dance experts has shown that a 'specific configural perceptual mechanism' that helps the perception of a familiar action configuration can be enhanced by both visual and motor familiarity with that specific action. However, expertise turned out not to make a difference for the perception of inverted stimuli, with inversion being rather uncommon in dance (Calvo-Merino, Ehrenberg et al. 2010). Below, we will learn more about the involvement of mirror neuron systems in both recognition of an action and in the simultaneous preparation of motor responses. Cf. reviews in (Casile, Caggiano et al. 2011 ; Iacoboni 2009). Mirror neuron activations during both observation and performance of action has suggested that they are responsible for shared motor representations, associated with the motor intentions under discussion or Searle's intentions-in-action (de Vignemont and Haggard 2008).

This flexibility and adaptivity stems from two aspects of the representations that are implied in the expert's sculpted space of actions and less so in novices. Before mentioning these two aspects, let us refer to our discussion of the process of representational redescription in chapter II.2, which takes place during development and learning (Clark and Karmiloff-Smith 1993; Cleeremans 1997 ; Karmiloff-Smith 1992 ; Mareschal, Johnson et al. 2007). For example, investigating children's drawings, Karmiloff-Smith observed that initially children would draw houses and figures in a rigid sequential order and with fixed components. With increasing experience, however, children demonstrate employment of redescribed and increasingly structured representations which grant them more flexibility and creativity in drawing. Being asked to draw a house or man 'that does not exist', for example, they had to retain the informational core of the object while modifying some important feature of it (Karmiloff-Smith 1990). Similar modification processes also affect the motor intentions that support experts' extraordinary performances.

First, the representations that experts' cognitive processes employ tend to have more structure than in novices, with hierarchy being a central feature of that structure. Second, this hierarchical structure is such that it leaves some room for variety, allowing experts the necessary flexibility and adaptivity. These aspects where added to the chunking theory as research showed that chess experts are also better than novices in recognizing and memorizing board positions that deviate somewhat from familiar positions, while still being not much better with random positions (Gobet 1998).[338] Apparently, representations tend not to be chunked or compressed in an indiscriminate or unstructured manner. Instead, apart from chunks that represent bits of complex and specific information experts also develop templates. Templates are more complex information representations with a hierarchical structure that contain both a core pattern of information and some free slots for fine details that can be filled in according to a specific situation. With this extension, the 'template theory' can both account for enhanced expert capabilities and for the flexibility that are associated with these could (Gobet and Simon 1996).[339] In addition, template theory can explain why

---

[338] The template theory can better account for expert's complex pattern recognition and response generation than the preceding chunking theory. As such it can counter some of the objections made by the Dreyfus's alternative account of intuition (Dreyfus 2004), more based upon phenomenology than associated with mechanistic explanation (Gobet and Chassy 2009).

[339] Acquisition of templates is similar to chunk acquisition, with templates being established when there is enough overlap in information between perceived patterns with some variety in components (open slots) that are associated to the overlapping pattern through particular similarity links (Gobet, Lane et al. 2001). Research of recall and problem solving with chess masters with different specializations demonstrated that their responses correlated with their having such highly  differentiated and complex knowledge structures, correlating with their domain of specialization and was not dependent on general mastery of chess (Bilalić, McLeod et al. 2009)

experts still perform better than novices in a kindred domain, even though they have little personal experience with it. For example, baseball experts can still benefit from their expertise when asked to memorize cricket events, presumably because several templates concerning these competitive ball team sports are valid in both sports (Jessup 2009). Moreover, even when experts actively perform in a domain in which they have no motor experience they appear to benefit from their expertise, transferring the assembled templates and having to only fill the slots in with the specifics of muscular information (Keele, Jennings et al. 1995)t.

In sum, this section has discussed how expertise is associated with the modification of motor intentions, affecting not just motor processes but also cognitive processes like perception, recognition and memory. This is in line with an account of theory of representations which are being shared between multiple processes involved in motor action and which are undergoing changes due to learning. Corresponding to these representational changes, we might expect the mechanisms underlying these motor intentions to change as well. It is to this aspect of motor intentions that we will now turn.

### 2.2.2 Motor intentions and mechanisms that change with growing expertise

In previous Parts we discussed various lines of evidence of the phenomenon that complex and dynamical systems tend to develop hierarchical (heterarchical) structures. Such developments are facilitated by the formation of generative entrenchments or kludges, as was discussed in Part II: the formation of a distinct component mechanism – often from already available components - as can be especially observed from functional differences or differences in performance of a cognitive function. Apart from the functional consequences of such kludge formation, we must usually also infer a change in the algorithmic theory which accounts for the performance, including the representations involved. It is to be expected that these processes of kludge formation and changes in representations can be observed also in the present context, where we are considering the interaction between a hierarchically - or rather: heterarchically - structured intentional cascade and the sculpted space of actions that is the result of growing expertise for a particular domain. What modifications obtain in with regard to the motor intentions that contribute to the effects of expertise?

Learning, memorizing and employing motor representations are complex processes carried out by complex mechanisms with components being distributed in the brain. Underlying every form of acquisition of expertise or skill learning are the various forms of neural plasticity that affect synaptic processes involved in interactions between

neurons. Based upon such modifiable neuronal interactions, learning a particular skill depends upon the crafting of cell assemblies, as is evident from modified motor representations in primary motor cortex upon relatively simple skill learning which last at least several days (Nudo, Milliken et al. 1996).[340] Hebbian learning allows further development of different types of such cell assemblies in the form of simple chunks and more complex templates, which are involved in complex and flexible motor skills as we learnt in the previous section (Chassy and Gobet 2011). The complex and distributed nature of the mechanism involved in such skill learning is evident from the study of apraxias and other deficits. For example, some patients cannot perform complex actions while being able to complete the simple components actions, others can perform complex actions as long as they are automatically initiated, still others have difficulty acquiring new action skills, and so on (Jeannerod 1997).[341] Such effects show how an agent's space of actions is sculpted partly under the influence of mechanism components that are difficult to explicitly control.

Indeed, skill learning does not always require complex mechanisms that involve higher levels of intentional control. Both computational and empirical studies have demonstrated that rather simple processes that derive statistical information from the observation and performance of actions do automatically lead over time to a hierarchically structured network, which is fit to represent appropriately the hierarchical structure of behavior and includes chunks of behavior (Botvinick 2008).[342] Nonetheless and in line with the expected mechanism modification, learning such complex and structured motor representations involved in motor intentions appears to be a complex process. Indeed, even in simple skill learning at least two phases can be distinguished which may even be considered as two different processes.

The phases to be discussed are to a large extent responsible for the two aspects that we distinguished earlier with regard to motor intentions: these involved both the recognition of affordances in the environments and the preparation of appropriate

---

[340] Research suggests that a change in motor representations is less dependent upon mere frequency of use but occurs in particular when a new motor skill is being learnt (Plautz, Milliken et al. 2000).

[341] Related to though different from the discussion here of motor intentions and expertise is the research in habits or habituated responses. Chunks of behavior are in that case automatically performed upon the perception of a particular stimulus even though the reward with which these habits earlier were associated may have become devalued (Graybiel 2008). It is the fixed nature of habits and the lack of competition between potential action responses in a given situation that allows distinguishing them from the more functional and adaptive forms of expertise discussed here, although that distinction may be gradual rather than strict.

[342] Several lines of evidence confirm that subjects typically encode observed actions in a hierarchical fashion. Indeed, the better they hierarchically encode behavior, the better they are capable of learning and reproducing it. Hierarchical encoding can partly rely on the observation of (statistical) changes in speed and direction of movements, which allows the first stage of segmentation of behavior (Hard, Lozano et al. 2006). Research suggests that infants but also adults use statistical processes for segmentation, which subsequently facilitates further understanding of intentional and hierarchical action (Baldwin, Andersson et al. 2008 ; Baldwin and Baird 2001).

motor programs. Interestingly, a first indication that expertise affects neural processing in several ways was obtained with fMRI investigations of verbal and motor expertise in which subjects would gain experience with tasks that required them to read and generate words or to complete maze tasks of varying complexity. Results have lead to the distinction of an early phase in which repeated behavior lead to an increasing efficiency of corresponding neural activations. This was followed by a second phase, characterized by additional activations in different neural areas, which has been associated with an increased and more comprehensive access to sets of stored associations or programs related to the initial task. Based upon these findings, the authors conclude that experts can be said to perform 'different tasks' than novices do (Petersen, van Mier et al. 1998): novices cannot rely on the structured representations and associated response options that make up an experts' sculpted space.

Concurring with the distinction of two aspects of expertise and specifying the finding of two phases of increasing expertise is the distinction of an early phase of improved pattern recognition and a subsequent - though partly parallel - phase in which response options are becoming associated with such patterns. Pattern recognition being reliant upon long term memory, these processes unwind largely automatically and therefore require hardly any short term memory (Chassy and Gobet 2011).[343] Otherwise put, during the first phase and thanks to the establishment of chunks a decrease in working memory involvement and consequently an increased efficiency can be observed. During the second phase – in which additional activations were found by Petersen and colleagues (Petersen, van Mier et al. 1998) -  a functional reorganization takes place, in which more complex and fine-tuned knowledge structures (chunks and also templates) from long term memory are activated in experts, in association with working memory contents of an ongoing task (Guida, Gobet et al. 2012).[344] In sum, kludges are established with increasing expertise, affecting expert performance in at least two ways which are observable even at the level of relatively simple motor intentions: these

---

[343] Focusing on the role of memory in expert performance, the LTWM framework explains this as based upon 'long-term working memory' in which long term memorized information is used in interaction with rapidly accessed short term memory (Ericsson and Roring 2007). The difference between this framework and template theory is perhaps less than stated, as the latter as well includes an interaction between long and short term memory (Gobet, Lane et al. 2001).

[344] Corresponding with such a transition is the observation of a shift occurring during succesfull learning of motor representations or schemas with activations tending to rely less on anterior frontal regions but instead more on posterior regions (Tracy, Flanders et al. 2003). Investigation of tennis players with different levels of expertise shows that expert players were better capable of grasping and long-term storing the hierarchical structure of the tennis serve, which correlated with the quality of their reproducing it (Schack and Mechsner 2006). Research with judo experts suggest that the study of memorized cognitive hierarchical representations of throwing techniques in individuals can enhance effects of their training these. The authors note that it is striking that experts across different sports use comparable hierarchical representations (Weigelt, Ahlmeyer et al. 2011). In music, experts are shown to equally have enhanced processing and memorizing of hierarchical structures which facilitate retrieval and practice of difficult parts (Williamon and Valentine 2002).

neural assemblies process information with greater efficiency and have undergone some functional reorganization as well that also affect motor responses.[345]

Now it may well be contended that the modifications of mechanisms that are involved in expertise are not unlike the shift that occurs in modes of processing according to the dual-process accounts that we've discussed in the previous Part.[346]In that context, we've noted that not only are representations modified but underlying mechanisms do not remain the same, as well. Indeed, dual-processing may be generally involved in expertise, as experts do employ both automatic processing after expert intuition has been formed, and also engage in controlled, deliberate search for solutions when necessary (Campitelli and Gobet 2010).[347] In a more restricted fashion, dual-processing theory can be applied to the chunking process itself, which is so important for the acquisition of expertise. In that case, we can distinguish between automatic and deliberate chunking processes (Gobet, Lane et al. 2001).[348] In the latter case, the characteristics of the more comprehensive templates and the chunks that are formed during learning are determined partly by deliberate choice – which confirms that indeed some top-down influence is effective, concurring with the intentional cascade framework. More below in this Part we will consider to what extent agents are capable of deliberately determining how to represent complex information and in doing so are also determining the corresponding neural processes. For now it suffices to point out that informational encapsulation of these chunking processes appears not to be at stake.[349]

---

[345] Such kludge formation and the corresponding increase of connectivity that is responsible for activations in related, associated areas can be explained in terms of modularity, as was suggested in the previous Part. Indeed, a recent computational study which compared the evolution of networks under conditions of selection pressures for both performance and connection costs demonstrated that modular networks fared better with regard to both conditions. The authors point out that when a system develops a modular structure this also implies that there are fewer parameters to optimize, fewer nodes to connect and hence smaller connection costs, smaller effects of mutations on the overall system – all contributing to fast and cost-effective adaptability (Clune, Mouret et al. 2013). Expertise, when considered in terms of kludge formation, yields the same benefits, so we would argue.

[346] Automatic processing, so we found, is considered to be intuitive, fast and parallel, involving implicit and non-verbal knowledge and requiring no conscious attention. Controlled processing, on the other hand, is considered to require consciousness, to be slow and sequential and to rely on explicit knowledge. Behavioral and computational differences are furthermore associated with different underlying mechanisms (cf. (Frankish and Evans 2009).

[347] Somewhat differently put is the distinction between intuitive and analytic processing inspired by dual-processing theories (Hodgkinson, Langan-Fox et al. 2008). These authors point out that emotion plays an important part in intuition. Others agree with the importance of emotion in expert intuition (Chassy and Gobet 2011), understanding emotion according to Frijda's notion of emotion as action tendency (Frijda 1986).

[348] Studies of music and sport experts show that deliberate practice has a strong effect on expert performance. Such experts deliberately practice specific components or features of their performance, like difficult runs on the piano or take-offs in sports or the tempo changes in the piece as a whole, thus further elaborating the hierarchy of the representations underlying their performance (Ericsson, Krampe et al. 1993 ; Meinz and Hambrick 2010).

In any case, we have argued in this section that corresponding to the modifications of the representations involved in motor intentions, underlying mechanisms are changing – for example due to kludge formation. The establishment and memorizing of complex and fine-tuned knowledge structures and their association with one or more appropriate response options leads to both increased efficiency and reconfiguration of the relevant processes and the mechanisms responsible for them (Guida, Gobet et al. 2012 ; Petersen, van Mier et al. 1998).[338] In sum, this repertory of knowledge structures and associated response options in a sculpted space of actions allows an agent usually to respond appropriately without requiring the agent to pass through the entire intentional cascade. Instead, he can rely on the expertise he has gathered over years of practice – practice which at the time did rely on the necessary distal and proximal intentions.

### 2.2.3 Motor intentions and differential generative entrenchment of components

From previous Parts we learnt how mechanism modifications underly changes in cognitive functioning. More specifically, we learnt about the associations between kludge formation and the changes or redescriptions of the representations involved in cognitive processes. In those contexts we also learnt that not all mechanisms components or representational components will equally be involved in future developments, as some are more generatively entrenched than others. When expert learning sculpts the expert's space of actions, therefore, we may expect that some dimensions or areas of that space will figure more prominently in further developments, than others. As we will see in this Part, such generative entrenchment obtains at all levels of the intentional cascade.

Generally, it has been observed that developing or learning new actions does often not imply the formation of completely new action representations but instead builds on previously learnt action representations, modifying these in some respects (Jeannerod, Arbib et al. 1995). In the present context of motor intentions, we did learn how experts demonstrate with their cognitive responses increased efficiency in processing and greater availability of appropriate responses after having established

---

[349] This distinction between automatic and deliberate chunking processes would explain why the template theory of expertise in itself does not contradict the fact that deliberate practice can contribute to acquisition of expertise, the latter being defended by (Ericsson and Roring 2007). Deliberate practice has been shown to be effective in many artistic and scientific and sports activities, among which chess (Ericsson, Roring et al. 2007).

[350] The basal ganglia are important for chunking action sequences as several studies show. Patients with lesions in these structures are impeded in developing novel chunks, contributing to their lagging behind in responses to repeated tasks compared to healthy subjects (Boyd, Edwards et al. 2009).

knowledge structures like chunks and templates (Guida, Gobet et al. 2012 ; Petersen, van Mier et al. 1998). These templates, we noticed, contain empty slots which allow them more flexibility than if they would have been completely specified. Granted with this flexibility and with the set of response options associated with them, they are applicable in a wide range of situations and thus likely to become more generatively entrenched (Gobet and Simon 1996).[351]

Such differential entrenchment of chunks and templates is confirmed by much research that focuses specifically on the influence of hierarchical structure in representations that are employed in learning and imitation. Obviously, this holds when subjects are explicitly articulating hierarchical relations in an object assembly task, in which case subjects who developed more elaborated hierarchical representations of action were not only better in understanding and recalling an action but also in performing it (Hard, Lozano et al. 2006). However, similar consequences of differential generative entrenchment have also been observed in studies that focused on implicit motor intentions and in non-human animals.

For example, young children automatically develop hierarchical representations of an action they are required to imitate. Indeed, the phenomenon of 'overimitation' demonstrates that they have a greater capacity to represent quite elaborate hierarchies than animals do even though language is not directly involved (Lyons, Damrosch et al. 2011).[352] More specifically demonstrating differential entrenchment of knowledge structures in animals is research in which greate apes demonstrate their capability of learning hierarchically structured actions for which they employ representations at both the level of action sequences and at a higher level of action 'programs'. These program level representations allow greater flexibility and can be applied in a wider range of situations, for example when a nettle leaf eating program can be further specified when a specific plant is targeted (Byrne and Russon 1998). Particularly the program level representations at stake here are comparable to templates and similarly allow greater flexibility and adaptivity as they allow further specification of components that are left open.

---

[351] To the extent that the brain prepares in parallel several options for action while interacting with the environment and internal - personal - conditions it is plausible that knowledge structures which bring along such properties will influence the 'saliency maps' associated with this parallel processing positively for more entrenched options (Cisek and Kalaska 2010). Several authors subscribe to this phenomenon of a competition between response options, e.g. (Brass and Haggard 2008 ; Botvinick 2001);

[352] The authors suggest that overimitation in children can be explained by referring to the 'teleological stance' that they ofen appear to take. This stance implies that - especially - children interpret action as aimed to realizing a particular goal rather than fulfilling a mental intention  (Gergely and Csibra 2003). Overimitation being not restricted to western cultures, it may be especially important for the transmission of complex forms of tool-use, which is as much prevalent in humans in contrast to animals as overimitation is (Nielsen and Tomaselli 2010).

This difference between levels of action representation and their differential entrenchment has also been demonstrated in human experts. For example, in studies of sports expertise, experts generally demonstrate a more differentiated hierarchical structure in their representations, which does contribute to their better performances. But even at the level of the 'basic action concepts' that occupy the bottom of these hierarchies researchers found that there was more correspondence between experts than between non-experts and novices, (Schack and Hackfort 2007 ; Schack and Mechsner 2006).[353] Such 'basic action concepts', however, are more difficult to transfer to another domain of expertise as throwing a ball in baseball and in cricket require motor actions that are quite different from each other. Nonetheless, as mentioned above, in some cases expertise can be transferred even to a different action modality, as when subjects had learned to perform a particular sequence with their fingers, which did facilitate sequence performance with their voice: apparently the sequence was represented as a template which could be filled in with the specifics of muscular information (Keele, Jennings et al. 1995).[354] Similarly, it has been found that sport expertise does rest upon the establishment of templates that can be employed by experts in a wide range of situations, even outside the domain in which they are specialized. Baseball and cricket experts, for example, were found to employ templates that contain representations of strategic positions of offensive and defensive players and their respective goals in each other's domains, even though they had no expertise there (Jessup 2009).

The advantages associated with this fact of differential entrenchment of action components will come to the fore once more when we will be discussing the proximal and distal intentions that an agent uses when planning his future and current actions more below. One of the main reasons will turn out to be the support that such action components offer to enhanced coherence and consistency in his actions. Let us take a short look at this characteristic before moving to the next level of the intentional cascade, to the proximal intentions.

---

[353] These 'basic action concepts' refer to observable - and commonly observed - components of complex actions, like throwing up a ball and hitting it in a tennis serve. The concept is remotely connected to the philosophical debate about the question whether there are basic actions - bodily, mentally, causally or otherwise (Annas 1977). Here is not the place to go into that discussion.

[354] This is for the authors reason to talk about the 'modularity of sequence representation' (Keele, Jennings et al. 1995). Although Pacherie also talks about the modularity of motor intentions, she does include in those not just the sequence representation but also effector related information about biomechanical constraints and kinematic and dynamic rules that govern the motor system (Pacherie 2008), which Keele et al. consider as belonging to the 'effector system'. This does not preclude the possibility that these motor intentions indeed can be further decomposed, with the sequence representation being a relatively separate component.

### 2.2.4 Motor intentions and consistency of action

Motor intentions were found to consist of representations that contain information about environmental affordances and motor movements pertinent to a certain action (Pacherie 2008). It seems plausible to assume that such complex representations, when established, allow an agent to act consistently in a recurring situation. This consistency may indeed depend upon the kind of guidance mechanism that Frankfurt referred to, lingering in the background during the unwinding of an action, waiting to intervene only in case of a deviance or loss of control. We found in section III.2.1. Frankfurt explicating that guidance implies: "a certain consistency or steadiness of behavior; and this presupposes some degree of persistence" (Frankfurt 1988 84).[355] Admittedly, this quote refers primarily to behavior at longer stretches of time but we've meanwhile argued that long-term practice also has an impact on short-term actions and on the on-line mechanisms that provide guidance for an ongoing action. To the extent that an agent has over time established a sculpted space of actions, we may expect him to act particularly consistently when actions that pertain to that space are being employed.

In fact, our previous section suggests that we may distinguish between action components at different hierarchical levels, or their representations. The templates that we discussed earlier figure at a higher hierarchical level than the knowledge structures that contain very specific motor representations, which may fill up the empty slots in those templates. Indeed, the differential generative entrenchment of action components and their representations is directly related to the consistency an agent demonstrates between his actions: the more a particular component is entrenched, the more we should expect to observe its functional properties in these actions. As a result, we should expect consistency in action to be distinguishable at different levels of specificity, as well.[356] However, given that we expect an expert to be flexible in adapting to specific environmental conditions and changes, consistency is especially supported if experts indeed employ templates with free slots that allow for adaptivity.

Consistency obtaining at different levels of specificity has been observed in experiments. For example, in research that required subjects to decide how they

---

[355] More below, we will discuss coherence in action and not just consistency. Though both terms indicate that an element satisfies a particular constraint or fits together with another element, coherence is usually taken to be more difficult to obtain as it is taken to involve a wider set of constraints than consistency (Thagard and Verbeurgt 1998). Even though it is debatable whether coherence is useful as a criterion to select scientific statements, its use in practical and ethical matters is more obvious as there is usually less consensus about which statements deserve support, or not(Millgram 2000).

[356] Approaching a comparable question from a dynamical theory perspective rather than from a mechanistic explanation perspective (which are said by the author to potentially complement each other), Kelso argues for the prevalence of 'synergies' in brain and behavior. Such synergies are structural and functional units that are 'soft-assembled' under certain conditions but gain stability over time and can then support coordination and control within an organism (Kelso 2009). Synergies can contribute to the theory of embodied cognition, positing that such soft-assembled units which span brain and body (and environment) can be employed in several functional domains (Anderson, Richardson et al. 2012, (Anderson, Richardson et al. 2012)

would play the ball after viewing a netball videoclip, experts and novices significantly differed in the correctness of their responses, demonstrating effects of expertise with a domain. However, within a sports domain it is possible to further distinguish positions. Indeed, expertise with certain field positions was reflected also in differences in correctness between the experts (Bruce, Farrow et al. 2012). Focusing on a still more specific movement, research with animals and humans has shown different levels of consistency that can obtain in movements for reaching and grasping a cup or other objects. Indeed, templates – called 'schemas' by the authors – were employed during reaching that did leave room for grip specification and even grip modification when object position was changing. Grip specifications were more specific in their motor representations and did allow less flexibility, testifying to their lower hierarchical position and limiting their involvement under multiple conditions (Jeannerod, Arbib et al. 1995). Consistency can also obtain with regard to a particular dimension of an action under different circumstances, as long as relevant representations are shared. For example, motor intention and motor imagery share largely the same representations, which is demonstrated by a remarkable consistency in the temporal structure of an action, irrespective of it being executed or imagined (Jeannerod 1994). Such specificity of consistency in motor intentions and its dependence on particular action representation components is also observable in patients who are incapable of verbalizing a grasping task and its representation and are correspondingly less capable of consistently performing it when a temporal delay between stimulus and response is given (Rossetti 1998). Apparently, consistency between actions does especially require knowledge structures like templates, that can be shared between different conditions as they allow further specification. Given the importance of objects and goals in many motor actions, it is relevant to note that action consistency is indeed supported by hierarchical representations of which different components can become activated relatively independently (Jeannerod, Arbib et al. 1995). The presence of different levels of specificity in this context implies that some levels of representation are more at a distance from direct motor control. This hierarchical structure of the action representations allows consistency of action to appear at several levels, as well.[357]

Generally, there is reason for an expert to rely within his domain of expertise on the motor intentions that he has established over time as research with expert handball players also shows. Given that these motor intentions are the result of much gathered experience, it is not surprising that the action option that experts first generate in response to a particular game position are usually better than options that are generated secondly or later (Johnson and Raab 2003).[358] Apparently, their motor intentions allow them indeed to perform consistently well, associating the recognition of complex

situations with appropriate responses. Moreover, the fast and implicit responses of an expert bring along the advantage that he is not distracted by the alternative solutions that conscious deliberation may suggest even in a situation where the constraints are such that there is little room for alternative solutions, as research of base ball catching demonstrates (Reed, McLeod et al. 2010).[357]

Such a difference between implicit processes, involved in motor intentions, and explicit processes does of course also lead to differences between the levels at which consistency between performances can be observed. Abstract representations become especially more generally available as templates for other modalities when an agent explicitly attends to these and grasps the abstract rule underlying a sequence, whereas a motor sequence can be learnt implicitly by merely repeatedly performing it and establishing relevant chunks as a result (Dominey, Lelekov et al. 1998).[360] Particularly at higher levels of the intentional cascade may we expect such abstract representations to play an important role. As a result, we may expect that when we are focusing on proximal and distal intentions, we will also observe other forms of consistency at those levels of action. More interesting, still, would it be if we could find interactions between different levels of the intentional cascade and influences of such interaction on the formation of kludges in the mechanisms that are responsible for cognitive processes and action. Before looking at those higher levels and such interaction, however, some neural evidence concerning motor intentions will be considered.

---

[357] Obviously, consistency in responding to a complex situation is dependent upon the number of constraints that play a role in an agent's capability of pattern recognition and the specificity of the action option generated in response (Johnson and Raab 2003). This is the reason why an expert will act more differentiated in response to subtly differentiated situations. Moreover, the fast and implicit responses of an expert bring along the advantage that he is not distracted by the alternative solutions that conscious deliberation may suggest even in a situation where the constraints are such that there is little room for alternative yet optimal solutions, as research of base ball catching demonstrates (Reed, McLeod et al. 2010)..

[358] From a meta-analysis of research on decision making in sports emerged that most consistent effects of expertise were visible when subjects were asked to respond behaviorally to a stimulus and not when responding verbally (Travassos, Araújo et al. 2013). This confirms our notion that motor intentions are implicit and can only in a limited sense be explicitly verbalized.

[359] However, there are some other modulating factors involved in constraining the presence of an abstract representation of an action and its potential modifiability. For example, in some conditions there is a dominant limb effect for a particular action, implying an asymmetry regarding the effector of the represented action. This may be a consequence of task complexity influencing the tempo in which a translation of visuo-spatial characteristics to motor characteristics is carried out by practicing subjects, which is necessary for task performance. In any case, it does contradict the assumption that always an abstract representation is available and supports the transfer of a task to a different effector or modality (Panzer, Krueger et al. 2009)

[360] This is confirmed by research of a key sequence pressing task where participants did not detect the similarity of sequences in different tasks and did not demonstrate in their reaction times in the separate tasks an automatic transfer. However, the author remarks that differences in instructions can also lead to different processing modes and thus to differences in the transfer (Verwey 2003). Indeed, also evidence from a different strand of research, on imitation, shows that preceding task instructions - to observe or imitate - yield different mirror neuron system activation patterns, suggesting that task instructions prime or activate different neural networks, with potentially different tasks being performed by the same systems (Vogt, Buccino et al. 2007)

### 2.2.5 Motor intentions and some evidence concerning their neural implementation

Part I contained discussions of both Marr's and the mechanistic explanatory approach to cognitive neuroscientific research. Both approaches concurred that mechanisms and mechanism components presented as explanations for cognitive functions require at least three different types of analysis, these being an analysis of the task or function at hand, an algorithmic analysis of it and finally considerations of the possible implementation in a physical system – neural or otherwise. Although Marr himself may have been less explicit about how these types of analysis should be related to each other, in fact his work spurred future researchers to engage more with each other's work (Kosslyn and Maljkovic 1990). According to the mechanistic explanatory approach, insights pertaining to different levels of analysis and to different levels of mechanism can be used as 'mutual constraints', thus reciprocally limiting the theoretical options available at those levels (Craver 2007).

With regard to the current context, we're interested in evidence about implementation of the motor intentions' capability to integrate multiple kinds of information. Moreover, the neural implementation of these motor intentions should be such that modifications obtain due to learning and expertise, such that generative entrenchment can occur. Such entrenchment should have lead, among other things, to involvement in ever more processes, compared to less prominent motor intentions. Most of the evidence concerning motor intentions referred to in the previous sections stem from developmental or cognitive studies, from animal studies and from computational studies. These studies have suggested that certain types of representations and processes are involved in motor intentions. For the elaboration of more detailed and comprehensive mechanistic explanations, further insights concerning underlying neural component processes and parts would be required. Given the limitations of our task, the evidence presented below should only confirm that in concurrence with the structured organization and modifiability of motor intentions, so are the underlying neural activities structured and modifiable.

Just like we observed that the impact of expertise on motor intentions can be specified at multiple levels, so can we find neural implementations of this at the level of single cells but also at a more comprehensive level. Evidence shows, for example, that experience and expertise does affect the activation patterns of single cells in correlation with specific actions. In sequential actions, for example, single cells in prefrontal cortex have been found to be activated in correlation with particular steps of an action sequence that a monkey must perform (Mushiake, Saito et al. 2006). Intriguingly, a specific cell type - mirror neurons - is activated both by observation and

by performance of complex actions like grasping an object and bringing it to the mouth. Adopting Jeannerod's idea of motor representations being shared between conditions, it was immediately assumed that these cells would undergird an observation/execution matching system by representing a 'motor vocabulary' (Gallese, Fadiga et al. 1996 ; Rizzolatti, Fadiga et al. 1996). Such shared representations have meanwhile been investigated via studies of mirror neurons and mirror neuron systems that are involved in the representation of a huge variety of stimuli and actions, observed in virtually all perceptual modalities and during the performance of simulation, verbal or motor tasks (see recent reviews (Casile, Caggiano et al. 2011 ; Cattaneo and Rizzolatti 2009 ; Glenberg 2011 ; Keysers and Gazzola 2009)).

Just like motor intentions were found to represent not only motor actions but also to include representations of relevant environmental affordances for action, so do these neurons (and even large neural networks) have the perceptual-motor qualities that facilitate such interactions (Casile, Caggiano et al. 2011). Indeed, such mirror neuron activation supports the brain's function of anticipating future action outcomes, making it more likely that these will be consistent with previous experiences (Kinsbourne and Jordan 2009).[361] These neuronal activities exemplify how a relatively simple and low level component mechanism may play a crucial role in facilitating an agent to interact consistently and coherently with his environment without continuously demanding attention and conscious cognitive processing.

Such interactions between environmental conditions and motor responses occur also at a higher level of specificity. Have mirror neurons been associated with quite specific motor actions – or component actions - in response to particular stimuli, more comprehensive action responses to particular environmental conditions rely on larger neural networks. Habitual responses or habits, considered as an action sequence in response to particular conditions, refer not just to chunks of behavior but demonstrate the kind of flexibility that fits with our notion of templates. Such habits develop over time and rest upon a network involving loops between cortical areas and the basal ganglia, with striatal neurons playing a role as throughput (Graybiel 1998). The stronger these cortico-basal ganglia loops become, the less flexible are the habits with regard to both stimulus and response conditions (Yin and Knowlton 2006). Indeed, obsessive-compulsive disorders that are associated with uncontrollable and inflexible actions in patients, have been associated with these same loops (Graybiel and Rauch 2000). In such patients, the motor intentions associated with these actions are no

---

[361] The motor representations that become activated upon the perception of an affordance can also include tool specific features when the subject has become familiar with that tool (Valyear, Gallivan et al. 2012).

longer integrated in the intentional cascade as other actions are.

For in most cases, chunked motor actions are still integrated in a larger neural network and consequently potentially affected by proximal and distal intentions. It is the prefrontal cortex that is held to be crucial for the integration of information that is involved in the various tasks that support the performance of action, like processing sensory information, associating information with motivational values and activating of necessary representations from memory (Forbes and Grafman 2010 ; Tanji and Hoshi 2001). Apart from a distinction between task-specific areas of PFC, hierarchical structures of the brain have been correlated with hierarchical structures in action representations.[362] It appears that the hierarchical structure of action, which we've argued brings several benefits along, is facilitated by a similarly hierarchical organization of its neural underpinning. Within the prefrontal lobe, motor memories are being stored along a gradient in terms of their complexity (Fuster 1997). Evidence shows how simple motor acts are being stored in the posterior premotor cortex, with anterior areas being more involved in complex forms of action control (Botvinick 2008).[363]

In sum, we can find in the brain relatively small mechanisms that are responsible for simple motor actions in response to particular stimuli and more comprehensive mechanisms that are associated with acquired motor sequences to complex environmental conditions. Remarkable is this fact that we find at several levels of specificity an integration of the perception and recognition of these conditions and the determination of a motor response – whether at the cellular level or at the level of neural networks. This reflects our previous observations of motor intentions at several levels of specificity, for example as chunks or more flexible templates of action responses. Although these contribute to an agent's sculpted space of action, other contributions are needed if more comprehensive action plans are to be integrated in this sculpted space. Consequently, although in many cases actions can be determined by relatively simple motor intentions and correlated with specific component mechanisms, proximal and distal intentions must often be involved as well. These intentions do not only modify an agent's motor intentions and their neural underpinnings indirectly via long-term

---

[362] Obviously, there are several reasons to question such mapping of hierarchies onto each other, which easily slides into a kind of neo-phrenology (Uttal 2001). Such reservations notwithstanding, hierarchical models are not just constructs as several lines of research robustly support their reality (Cohen 2000).

[363] Botvinick and others have argued that hierarchical reinforcement learning can account for computational and empirical results of emerging hierarchy in action (Botvinick, Niv et al. 2009). Whether this model can do without the explicit representation of goals and without the model having a hierarchical structure itself has been questioned in (Cooper and Shallice 2006). Irrespective of this difference, these authors agree wth respect to the prevalence of hierarchical structure in action and the fact that some action components are more prevalent than others.

processes of learning and exercise. On top of that, proximal and distal intentions often interact with the representations involved in motor intentions for the determination of an action, since motor intentions do not only receive top-down influence but in turn exercise bottom-up influences.[364]

---

[364] Motor intentions are different from the simple stimulus-response couplings that can be found to underly habits which are also depedent upon a subcortical loop (Graybiel 2008). The response patterns of motor intentions are more complex and they are capable of being prepared by relevant distal intentions, which do not need to be continuously activated. Moreover, regarding the mirror neuron systems, evidence shows that not all perceptual information is immediately relayed to mirror neurons but some gates or filters appear to - sometimes implicitly - modulate that information, as we've discussed in (Keestra 2012).

# 3  PROXIMAL INTENTIONS: A MEDIATING ROLE

Above, in chapter III.2, we pointed out how amazingly controlled and intentional an expert singer performs his Don Giovanni, even though the complexity and comprehensiveness of his acting, singing, observing and interacting is such that we cannot expect him to deliberate and rationally decide about all of it on the spot. Instead, so we argued, his behavior is facilitated by a sculpted space of actions partly determined by implicit motor intentions, knowledge structures consisting of representations of complex patterns and associated motor movements, assembled over time with practicing, learning and experience. However, if only motor intentions would determine this sculpted space of action, this would raise several questions.

First, one could ask whether a sculpted space, determined by motor intentions alone, would facilitate the agent's task of selecting a single action option from the many motor intentions that a situation with many affordances may activate. Are there constraints available that help him to choose between those options, taking into account that motor intentions operate at a temporal scale that makes it impossible for deliberate and conscious choice to intervene? Constraints that are intimately related to his previous choices and experiences and can therefore be considered to be in line with these?[365]

Second, even if there are no alternative action options available, will a motor intention automatically ensue into an action given certain affordances, similar to those automatized actions that escape any form of control? Although Aristotle already argued in favor of habits as part of an agent's moral behavior, there are situations imaginable in which these habits are morally inappropriate. Therefore the question rises whether it is possible to block a motor intention's application under certain circumstances?

Third, apart from blocking a motor intention in an exceptional situation, are there perhaps further constraints on an agent's sculpted space of actions? For the absence of further constraints on motor intentions could lead an agent to act inconsistently over time or in ways that do not cohere with his other intentions or beliefs. Indeed, can we consider those other intentions and beliefs to contribute to an agent's personality in such a way that they further constrain his motor intentions as these on their own probably cannot constrain each other adequately?

For example, consider our expert singer who has to avoid confusion in an on-

---

[365] The role of the agent's history has become – again – important in the philosophical debate about his autonomy of action. With such history being given its due, autonomy becomes a more complex issue as former choices and experiences may contribute to ongoing and future actions even if such contributions are not always consciously deliberated.

stage situation with many stage props and persons that could offer various action options, pertaining to both his Saint François and Don Giovanni roles. How does he decide to which affordances he should respond and which intention - out of many intentions - to act he should implement in a given situation? Moreover, consider the situation of an opera star who is flown in for just one guest appearance in a regular series of performances of Don Giovanni. Being familiar with the score and libretto, he must still find himself at home in a new direction and stage, and so on. Under an exceptional director like Peter Sellars, who invited as singers for Don Giovanni and his servant Leporello twin brothers and set them as drugs dealers in Harlem, he may have to block some habitual actions while substituting them for newly learnt actions. Of course, he must prepare his role by looking at pictures of the set design, reading stage directions and formulating for himself intentions to act in certain ways or to avoid certain actions – depending upon his expertise with the role and with other directions. Having prepared himself in this way, the experienced singer may need only a short rehearsal of some crucial scenes to practice the performance satisfactorily. This rehearsal would allow him to put additional constraints on the set of motor intentions that his preparations alone could probably not do.

Clearly, distal intentions are involved in the singer's preparation by looking and reading and planning, yet proximal intentions must also be established, for example by rehearsing. These additional intentions further constrain and sculpt the agent's space of actions. With motor intentions being processed in a semi-modular way and being non-conceptually and unconsciously responsible for the transformation of specific perceptual information to corresponding sensory-motor information in a given situation, distal and proximal intentions offer further constraints. With regard to proximal intentions, Pacherie writes that, the: "problem at the level of P-intentions consists in integrating conceptual information about intended action inherited from the D-intention with perceptual information about the current situation and memory information about one's motor repertoire to yield a more definite representation of the action to be performed" (Pacherie 2008 185). Integration of these sources of information leads to the 'anchoring' of the distal intention in a given situation, which is an important task of proximal intentions.

Is this task of specifying the more abstract distal intentions for their eventual initiation and performance alone a crucial task, proximal intentions are allegedly also involved in monitoring and guiding the outcome of the action. These tasks imply the interaction between intentional contents that are partly hierarchically related to each other, and processes that may be differently structured, being constrained by those contents and by environmental and motor conditions. Execution of these tasks

requires processes that probably are quite different yet still have to be related to each other. Next to the processes corresponding to motor intentions, proximal intentions contribute to the determination of action as well. Since the time scale of proximal intentions is not so restricted as is the case in motor intentions, explicit and conscious perception and cognition are able to play a role, according to this framework. As a result, rational constraints in the sense of coherence and consistency constraints are at work at this level, too, with proximal intentions being responsible for keeping track of an optimal action performance and for controlling for potential side effects (Pacherie 2006 ; Pacherie 2008) – things that motor intentions are not capable of.

In as much as proximal intentions play an important role with regard to all the issues raised above, they contribute in important ways to the sculpted space we're investigating. As noted before, they don't do so in separation, as they meanwhile integrate also contents from higher and lower level intentions. So now we will first provide a philosophical analysis of this complex task and an argument for its importance, after which we will be turning to empirical evidence.

### 3.1  A philosophical analysis of proximal intentions

We started the sections on motor intentions above with a philosophical analysis of motor intentions, largely leaning on Frankfurt's account of guidance. Now that we're shifting our focus on proximal intentions and their intermediate position, it will be Bratman's conception of present-directed intentions that guides our analysis, partly because Pacherie has explicitly built her framework on this conception.[366] However, it has been Frankfurt's introduction of the distinction between different orders or levels of intentions into his account of action that has influenced both (Frankfurt 1971). Frankfurt has argued for these levels of intentions and their interaction as a structure that is required to regulate an agent's actions. Moreover, he argues, it is this structured interaction between an agent's intentions that reflect his identity as a person (Frankfurt 1988). Bratman has taken up important elements from this account of intentionality by way of a structured interaction but has elaborated upon its dynamics and has developed a philosophical analysis that will turn out to be quite suitable for integration in an explanatory account of action. This is due to a large extent to his greater emphasis on temporal dynamics of the structured interaction (Bratman 1987).

---

[366] Proximal and distal intentions are terms that are being used by Mele, with proximal intentions referring to intentions for the 'specious present'. Bratman introduced the notions of future-directed and present-directed intentions for comparable purposes as proximal and distal intentions. Mele relativizes differences between the positions and simply notes in that context that 'terminology varies' (Mele and Moser 1994 65, footnote 10).

With that addition, the philosophical analysis has more to offer to an account like ours, which aims to contribute to a mechanistic explanation of action determination. Given that for our account we have a special interest in the dynamics involved and in the possibility of an emerging sculpted space of actions, alleviating in turn some of the tasks involved in determining the agent's appropriate action, Bratman's work will be more closely discussed in what follows.

### 3.1.1  Resolving conflicts between action options

In section III.2.1., we found Frankfurt analyzing guidance in terms of ongoing action monitoring and intervention mechanisms that remain passive as long as their intervention is unnecessary (Frankfurt 1988 75). In the current context, we find him interested in the different degrees of commitment or engagement an agent can enjoy regarding his decision to perform any action at all. In both cases, gradual variety obtains for an internal process that is relevant for an agent's actions.

Appreciating the fact that in any given situation an agent may have the troublesome experience of having multiple desires to perform certain actions, Frankfurt has introduced the notion of different orders of desire or volition. An agent may in such a case not be content with the desire that has eventually won out the competition, in which case: "he wants to be motivated effectively, with respect to the alternatives he faces, by some desire other than the one that actually moves him to act as he does" (Frankfurt 1988 48). The agent's internal conflict can therefore be twofold.[367] First, the conflict may be among his multiple first-order desires, each of which concerns a particular option to act.

Second, another internal conflict would occur when a first-order desire has eventually won out the competition and would determine the agent's action but turns out to be in conflict with a second-order volition that he also has. In that case the latter volition would have preferred the action to be determined by an alternative desire. The question is what role these conflicts and their solutions play in a given situation where several action options are at stake and where an agent can only execute a single action. Are there strategies or constraints available that facilitate solving these recurrent problems? Is an agent always forced to engage in a sequence of choices, in which he endures and solves the first type of conflict and then subsequently deals with the second type? Or are there other forms of interaction between the different orders

---

[367] An important feature of Frankfurt's account is that it is not uncommon for an agent that: "no second-order volition plays a role in the economy of his desires" (Frankfurt 1988 50). The phrase 'economy of desires' alludes to Frankfurt's view that in many cases, rational deliberation is not really at stake, but rather a competition between first-order desires in terms of the costs and benefits involved in their realization.

possible that would alleviate these tasks in some way or another?[368]

In a situation where different desires to act cannot be realized simultaneously, we are usually required to reject all except one in order to resolve this conflict. A simple solution to reach this result would be by ordering the desires and establishing a preferential ranking of action options, being members of the same order. These options then differ in degree rather than in kind for the agent. However, in a given situation it may be quite a large cognitive task to articulate and then order all action options, provided they would all be comparable to each other. Imposing a constraint that divides these options into different categories could be helpful in constraining this task. This is the case when an agent has specified a set of second-order desires and then identifies only with those action options that conform to these. Conversely, he may have decided that a particular desire should not belong to his particular space of action options and: "is finally to be excluded from the order of candidates for satisfaction" (Frankfurt 1988 68). The latter case refers to a situation in which the agent has put the rejected desires external to himself as a person and in which it would be profitable for him to sculpt his space of actions correspondingly.[369]

This analysis does not comprehensively show how an agent goes about to select and realize – anchor - one of his desires or intentions in a given situation but it does contribute to answering that question. Although Frankfurt focuses especially on the issue of an agent's responsibility for his actions, his analysis does also imply an interest in the contribution of the enduring structure of the agent's personality and identity to solving this question.[370] In his analysis he aims to demonstrate how this structure does constrain the potentially large cognitive task of selecting these action options among competitors and then realizing them. By deciding about the order of his desires and

---

[368] This does not mean that second-order volitions do not influence the competition between first-order intentions. Indeed, when Frankfurt later analyzes the wholeheartedness with which an agent can embrace his own beliefs or attitudes or intentions, he contends that the agent's satisfaction is with: "these psychic elements (…) rather than others that inherently (i.e. non-contingently) conflict with them, should be among the causes and considerations that determine his cognitive, affective, attitudinal, and behavioral processes" (Frankfurt 1999 103). This satisfaction implies that the agent will not resist or reconsider the outcome of the first type of conflict, between first-order intentions.

[369] It has been argued critically that Frankfurt eschews to assign a central role to objective or rational criteria for such decisions, which raises doubts about both the role of rationality in such an agent and about the moral value of them (Buss 2002). Frankfurt does indeed doubt about the nature of these decisions, as he explicitly admits (Frankfurt 1988 68). Moreover, in his reply to Buss he contends that it is well conceivable that living an immoral life may be good from the person's perspective, even if it is Hitler – so his decisions may be rational yet immoral, indeed (Frankfurt 2002b).

[370] Although as far as we are aware of, they do nowhere explicitly discuss each other's work, there is some affinity between this investment in the person as a final cornerstone of analytical philosophy of action and Ricoeur's phenomenological and hermeneutical philosophy of action. Indeed, the latter explicitly acknowledges that "the *question of personal identity* is posed at the point of intersection between the two philosophical traditions" (Ricoeur 1992 17, italics in original).

particularly by including and excluding some of these intentions from the space of actions that he is satisfied with, the agent constrains and guides these tasks: "[o]ne thing a deliberate decision accomplishes, when it creates an intention, is to establish a constraint by which other preferences and decisions are to be guided" (Frankfurt 1988 175). Some decisions, therefore, bear not just on a specific situation but have consequences for future situations as they can partly determine the processes that contribute to those future decisions.[371]

The upshot of this analysis is that an agent can try to resist "doing what comes naturally" by putting constraints on those processes that would otherwise determine in an unreflected manner his actions and thus sculpt the space of actions that remain open for him to do in a given situation.[372] The functions of these constraints are twofold: they create or contribute to the coherence of an agent's intentional actions and they provide a 'reflexive or hierarchical structure' between his desires and consequently to his identity, which does also contribute to the coherence of his actions. Indeed, Frankfurt applies elsewhere the notion of 'person' to this constellation. A person, he writes, is characterized by taking upon himself constraints that are not just limiting his thought and language, but also the "choices he can make" (Frankfurt 1999 113).[373]In sum, the process of the person's engagement with his own 'psychic characteristics' is comparable to what we refer to as the agent's sculpting process with respect to his space of actions: in both cases constraints on the available options are established such that a more or less coherent pattern of performances emerges.[374]

---

[371] Indeed, an assumption of Frankfurt appears to be that our decisions somehow influence the neural and cognitive processes that are involved in future decision making. He does not explicitly address this answer in the contexts discussed here, but others do. Taking up Frankfurt's influential notion of guidance control – discussed above in section III.2.1 - Fischer and Ravizza argue that in fact also the mechanisms underlying our decision making must be moderately reasons-responsive (Fischer and Ravizza 1998). Contributing to this discussion, Frankfurt articulates a very loose sense of 'reason' when he even ascribes reasons – but not beliefs - to insects that try to escape a predator (Frankfurt 2002a). Fischer agrees with that loose sense, writing that: "[a]n organism – any organism – can have reasons insofar as he or she can have interests or a "stake" in something" (Fischer 2004 149).

[372] Frankfurt has a keen interest in the psychological plausability of what a philosophical analysis suggests that an agent should mentally accomplish. For example, he notes that there may be a trade-off between the size of an agent's enlarged space of actions and his sense of identity: "[t]he task of evaluating and ranking a considerably enlarged number of alternatives may be too much for him; it may overload his capacity to make decisions firmly grounded in a steady appreciation of what he really values and desires" (Frankfurt 1999 109). Having established a stable set of constraints does, from this perspective, not so much restrict an agent's identity in a negative sense but rather supports it.

[373] With the importance attached to the person and its structure being determined by the constraints he has taken upon himself, Frankfurt's account offers little room for agents who try to justify an action retrospectively by adjusting their second-order volitions such that their action now complies with those. Such an agent, we would argue, has not determined the structure of his will and can be compared to the 'wanton' who is not moved by his will either (Frankfurt 1971).

[374] In an earlier text, his focus was less on the notion of the person but rather more on guidance and control. In that context, too, the emphasis is on the fact that an agent can avoid the performance of undesirable proximal intentions by constraining these, that is to: "replace the liberty of anarchic impulsive behavior with the autonomy of being under his own control" (Frankfurt 1988 175).

However, what remains to be shown is: why should an agent develop these characteristics at all, what good are these constraints? Indeed, Bratman has argued that a purely hierarchical account might be subject to a threat of circularity, it being that an agent might support some of his intentions to act by referring to his identity as a person, even though this identity is found to partly consist of precisely such selected intentions. Adding a temporal or dynamical component to this hierarchical account – since an agent is not a 'time-slice agent' -  it can be pointed out that an agent cannot but develop higher-order policies that allow him to self-govern his actions. Only by doing so can he at least try to avoid undesirable incoherencies and inconsistencies in his actions, for example by blocking an action in an exceptional situation, as will be discussed in the next section.[375] Nonetheless, these policies are not just based upon mere instrumental means-end reasoning but involve the agent's setting an end to his reasoning that is not instrumental in nature (Bratman 2002). Because of the constraints of his temporally extended agency, an agent must develop the constraints mentioned earlier – both contributing to the sculpting process as a whole. In what follows, we will further consider the constraints that pertain to the level of proximal intentions and – subsequently - their implementations.

### 3.1.2  Proximal intentions and blocking habitual action

As noted earlier, intentions occupy multiple positions in the agent's overall psychology with regard to his actions, ranging from prompters to terminators of deliberation, and from initiators to guidances of action. Corresponding with these positions are different functional properties, corresponding with the roles these intentions play in the complex and dynamic processes involved in his agency. An important role of Bratman's present-directed intentions – to which we refer as proximal intentions – is their being directed at the present or proximal situation, recognizing it as an appropriate situation for carrying out a specific intention: "[t]o have a present-directed intention to A, I must see that *now* is the time for action" (Bratman 1987 182, note 8, italics in original). Seeing that the present situation enables carrying out an intention is often, as we have learnt above, a matter of pattern recognition for which motor intentions are responsible. However, the recognition of a suitable affordance that is associated with a particular motor representation may trigger behavior that is not intentional but merely a habit – it may indeed even be contrary to an agent's current

---

[375] Obviously, with distal intentions in place, an agent is still not invulnerable for inconsistencies. First, it is improbable that an agent can specify all proximal – let alone motor – intentions to such an extent that he will never realize that in a given situation not all distal intentions can consistently be carried further. Second, as distal intentions often need time to be articulated and further specified, potential inconsistencies to be solved often emerge only at a later point.

intentions, at whatever level of specificity.[376] So how are the two different and what different cognitive processes are involved?

Bratman recognizes this risk of lapsing into habitual action in a situation where the outcome of that action would be at odds with our long-term or distal intention. Appreciating, like Frankfurt did, the limitations in cognitive resources and time for an agent to consider all available action options in a given situation, still Bratman draws attention to the important fact that an agent may at times be forced to 'block' the application of a general intention. He mentions the example of an emergency situation, in which case a car driver is forced not to buckle up, even though it does not mean that he reconsiders or abandons the general intention to always buckle up (Bratman 1987 88-89). The difference between a habitual response to a circumstantial trigger and an intentional response in Bratman's sense, is precisely this defeasibility of the latter as can be observed when it is being blocked in extraordinary cases. In such cases we can observe that there are several constraints or rules at work, interacting with each other in rather complex ways.

Proximal intentions are in a way a focal point of such interaction of multiple constraints. For Bratman has assigned to proximal intentions several functional roles. Proximal intentions play a role in initiating an action at a particular moment in a given situation, without depending upon a careful deliberation of the pro's and contra's of a particular action since that deliberation has been part of the distal intention against the background of which a proximal intention is formed.[377] However, anchoring a distal intention by way of a proximal action in a given situation does often require adjustments of the latter, without adjustments of the former. For example, adjusting the proximal intention to a changing situation involves 'temporal updating' it so that the 'now' remains appropriate (Bratman 1987 56).

Such updating of a proximal intention in order to anchor a distal intention appropriately or to block its application if necessary is influenced by some further constraints to which we will turn now.

---

[376] In footnote 329, we've referred to the fact that habits have a fixed nature and do not result from a competition between potential action responses in a given situation. See (Graybiel 2008) for further clarification of habits.

[377] Bratman refers to the 'hybrid character' that many intentions and action have with respect to their being deliberative, since many are only deliberative because of their being part of a comprehensive plan which is the result of deliberation and is itself not reconsidered in a given situation (Bratman 1987 30). Similarly, the standards or criteria to which such intentions and action should conform are only derived from such a plan or distal intention.

### 3.1.3  Proximal intentions and constraints for anchoring an action

Not just prompting but also exercising some control on an agent's temporally extended action, the question is whether particular standards or criteria are involved that may bring the agent to abort or instead release his action. Bratman argues that indeed such critera are in place and these are the criteria that we also apply when judging the rationality of an agent's explanation of his action in terms of his intentions. Such judgments are not only made regarding the distal intentions of an agent but also with regard to his proximal intentions. This argument results in the formulation of an 'intention-action principle': "the present-directed intention to A and the resulting action of intentionally A-ing are too tightly connected for us to praise the agent as rational for the former and yet not praise her as rational for the latter. This is because the intention and action are not separately controlled by the agent, but rather the agent's control of her action goes by way of her intention" (Bratman 1987 55). What standards or criteria are applied for this control process?

Bratman presents us with two different norms or constraints that – explicitly or implicitly – should apply to an agent's planning, if he is to optimally fulfill his intentions and plans.[378] These constraints respond to the fact that an action is never instantaneous nor isolated but interacting with other actions, intentions and beliefs – in ways, however, that are not completely transparent to the agent. The coordinating role of the agent's intentions and plans consist partly in putting constraints in place on the processes that eventuate in action and this holds not just for distal intentions but even for proximal intentions. The first constraint demands that a plan (or plans) should be not self-refuting but consistent, while the second constraint demands for means-end coherence of the plan. A planning agent is one who does take these two constraints into account, one way or another: "The recognition of these demands helps distinguish intentions and plans, on the one hand, from ordinary desires and valuations, on the other" (Bratman 1987 32). Given this analysis, an articulation of the constraints is relevant.[379]

Take the twofold consistency of an action plan, needed for effectively carrying out an intention. The demand for an intention's internal consistency is obvious as it is inconsistent if a singer plans to be silent and to sing a line simultaneously. Such

[378] Quite consistently, Bratman formulated these demands, constraints or standards  in his early 1981 article on means-ends reasoning (Bratman 1981), in his 1987 book (Bratman 1987), in (Bratman 1992a), his 2006 book (Bratman 2006b) and in his (Bratman 2009b).

[379] So it is foremost due to functional rather than moral considerations that we can us expect this complexity of the agent's cognitive processes with regard to proximal intentions. Indeed, Bratman argues that in the end the functionality of these proceses derives largely from "our interest in getting what we want" (Bratman 1981 262).

consistency forces an agent to also take his beliefs into account (while assuming that these are true) as he is constructing a representation of reality: "my plans should fit together with my beliefs into a consistent conception of the future" (Bratman 1981 259). Such a construction is difficult as it is complicated to account for the future since this is partly shaped by the agent's own intentions and plans – even in the case of a proximal intention and the immediate future that may result from it.[380]

The second constraint an agent should fulfill is a result of the fact that an action is usually not only temporally extended but also involves a hierarchy of steps or means that contribute to performing it.[381] Correspondingly, an agent's action plan needs to consist also of "subplans concerning means, preliminary steps, and relatively specific course of action, subplans at least as extensive as I believe are now required to do what I plan" (Bratman 1987 31). The specification of action, which the framework of the intentional cascade ascribes to proximal intentions, is at stake here. To begin with, it requires the agent to develop such a specification at all. He cannot pause with the mere formation of an action goal but should indeed proceed to specify means to realize the action – for example specifying whether to use precision grip or power grip while picking up a cross or sword. Moreover, these means should be included in his intention or plan if it is to count as such: we would consider it irrational for an agent to intend to reach a certain end yet not to intend executing a means which he believes to be necessary for reaching it (Bratman 2009b). Irrationality would in this case amount to dysfunctionality since an intention can not be realized without the intention of realizing its means. Determining the level of detail of this specification of the means, however, is difficult.

Clearly, it is not necessary for an agent to – perhaps implicitly – take every possible future situation in consideration, as some of these situations are implausible or unlikely to happen. Similarly, he is not required to specify in advance all minor

---

[380] Kolodny discusses the 'predictive significance' of intention (a term introduced by Scanlon) with respect to the future in connection with the constraints on practical reasoning. Kolodny underlines that consistency demands can make a rational agent to develop further intentions if these are related to actions that will facilitate the satisfaction of an earlier intention, which is different from mere means-end reasoning (Kolodny 2008).

[381] Bratman underlines that a planning agent is not a time-slice being, partially because his agency is extended in time (Bratman 2006b). If the agent is to self-govern, then he needs to realize that self-governance is not a time-slice phenomenon, too. Instead, it requires temporally extended planning with an important role for means-end coherence and limited room for reconsideration of his plans (Bratman 2007).

[382] Obviously, details of the muscle movements that are involved are even impossible to specify. Although most authors agree about this, there is still some debate about the question whether experts can or cannot articulate and verbally express details of their expertise. We've touched upon that debate in the previous Part. There we did i.a. refer to the implict and explicit stages of learning and development presented in (Karmiloff-Smith 1992), which presents a 'representational redescription' account of empirical research of expertise, including a final stage in which expertise can be made explicit. This account was contrasted with an analytical one that denied such explicitability of expertise (Dreyfus 2004).

details that will be involved in performing the action.[382] Depending on the agent's expertise, habits and skills, he must specify his intention more into detail or he can leave it up to the moment of performance and then rely on these for appropriately performing the action (Bratman 1987). For example, for a beginner without such expertise and habits, it is necessary to specify more into detail how he is going to perform next his canzonetta for Donna Elvira's maid, while an expert may trust his already stored proximal intentions to include the necessary specifications required at the time. This expert, though, may have to block and re-adjust his proximal intentions in a new and divergent stage direction. Clearly, anchoring his performance in that situation consistently and coherently may require him to specify his behavior, singing and instrument use completely different – not just during his preparations but also on-stage during his performance. A continuous complex interaction of established and newly formed intentions is the result, pointing towards the peculiar position of proximate intentions.

### 3.1.4  Proximal intentions and their peculiar position in the agent's psychology

The role played by an agent's expertise, habits and skills – which were also associated with motor intentions earlier - confirms Pacherie's observation that proximal intentions occupy an intermediate position in the framework (Pacherie 2008). Although proximal intentions can provide some amount of conscious guidance to the ongoing action, they are assisted in carrying out this complex task by the presence of motor intentions that can to some extent relieve it (Pacherie 2006). As a result, proximal intentions inherit constraints of a sculpted space of actions from both motor and distal intentions.[383]That they still have a specific role of their own is most obvious in a situation in which the agent demonstrates the 'defeasibility of general policies' by not applying a policy to the particular case, for example when otherwise a breach of some of his other constraints of consistency and coherence might occur (Bratman 1987). Our expert singer, for example, must be able during a extraordinary Don Giovanni stage direction to inhibit commonly juvenile behavior as his distal intention of complying with a director requires him now to act otherwise.

Let us try to shed some light on this peculiar position of proximal intentions. In some sense they appear to function like kludges that have been established in a responsible mechanism. Yet in another sense and unlike kludges, proximal intentions

---

[383] Hobson explicitly doubts whether Bratman's present-directed (proximal) intentions as such exist. Although he is correctly pointing out the priority of future-directed (distal) intentions, he overlooks the importance of the function of proximal intentions in anchoring the latter in specific situations and – at times - blocking their application (Holton 1999 246).

are also governed by constraints that are associated with conscious deliberation. What kind of explanation can be given for this peculiar nature of proximal intentions? To be sure, intentions and plans do not fulfill their roles in isolation from other components in the psychology of an agent.[384] Irrespective of the fact that we can subject them to those rational constraints as were presented above, they are also related to constraints that depend upon the physical and psychological beings that agents are. Indeed, the rational constraints themselves are partly due to our physical and psychological properties and limitations, which disallow us to perform two contradictory actions simultaneously, for example. Philosophical analysis aims partly to clarify precisely the interdependence that exists between our psychological structures – studied in psychology and neuroscience - and the structures of our thought and action (Bratman 2009a).[385]

Indeed, the present analysis of the process of sculpting a space of action aims to contribute to such a clarification, as well. It more specifically focuses on the properties of this sculpting process, which are partly dependent upon the constraints that have their source in practical deliberations that do at times determine the actions an agent performs. For another part, the constraints stem from the embodied and cognitive structures that constitute the agent and which also constrain the algorithmic and neural implementations of these deliberations.[386] Integral to these various constraints is the requirement that the space of actions does not remain completely fluid but gains a profile that is rather stable. By adding generatively entrenched properties to such a

---

[384] Intentions can be recognized as distinct psychological elements in the philosophy of mind, making their reason-giving status in normative philosophy only derivative, Bratman argues (Bratman 1981 263). There is some analogy in this view to Frankfurt's admission that it may be possible that an immoral life can be valuable and desirable for the person that lives it, who may not have convincing reasons to change it as long as it is coherent and consistent. Frankfurt explains: "the value to Hitler of living the life he chose would have been damaged by the immorality of that life only if morality was something that Hitler actually cared about, or if the immorality of his life somehow had a damaging effect on other matters that he cared about" (Frankfurt 2002b 248).

[385] Early on, Bratman committed to a functionalist approach according to which inputs are connected through psychological processes and activities to outputs like actions. There are regularities involved in these processes and activities, for which intentions, beliefs and desires can be responsible. An intention is then to be understood as a 'distinctive attitude' (Bratman 1987). In line with this account, Bratman much later acknowledged in his contribution to a symposium on consciousness that a specific intention needs not to be engaged in conscious thinking continuously. He referred to the Freudian idea of unconscious intentions that still have specific content, like 'sleep with your mother'. Consciousness would than be a higher-order – relational - phenomenon, that may or may not be at stake with regard to a specific intention at a specific moment (Bratman 2006a).

[386] As noted in Part I, such a multicausal account of human action is not new. Aristotle has offered a first account along such lines, famously contending that: "Thus every action must be due to one or other of seven causes: chance, nature, compulsion, habit, reasoning, anger, or appetite" (Rhetorics 1369 a 5-6). Our scrutiny of the role of a sculpted space of actions is partly inspired by the emphasis that Aristotle put upon the role of habit in action – including moral action in his Ethics.

sculpted space the agent will be faster and more flexible in his responsive actions, with decreased recruitment of resources in their performance. For the execution of his own complex actions as much as for joint or collective actions of a certain complexity, this sculpted space is quite advantageous (Keestra 2012).

Moreover, an important lesson that must be drawn from the consideration of this interdependency and the constraints that are associated with it is that an agent should basically not modify continuously his intentions and plans but should remain largely committed to them. Would an agent instead constantly reconsider and modify his plans, he would be seriously impeded in completing an intentional action at all (Bratman 2006b). This is partly so because generally a plan has ramifications and is temporally extended, rendering it likely that a modification will turn out to be counterproductive as it runs against earlier phases of its execution.[387] Apart from this argument about the irrationality of continuous reconsideration of one's intentions, Bratman mentions another reason why such reconsideration is disadvantageous to an agent – yet this reason is of a naturalistic nature and testifies to his ambition of developing an account of agency that fulfills several roles.[388]

If an agent aims to avoid counterproductivity as a result of his reconsidering his intentions, he is forced to carefully scrutinize the modification and its consequences which "is an activity that uses up time and other limited resources; while engaged in reconsideration I am unable to do other valuable things" (Bratman 1992a 6).[389]Apparently, this costliness has to be taken into consideration as well as it also affects the agent's adequate actions. When modifying his plans and intentions, he must also devote cognitive resources to updating his proximal intentions, impeding his ongoing performance.[390] In contrast to investing these resources in case of

---

[387] Given the temporal extendedness of any action, it is unsurprising that a methodological priority is given to distal intentions, as they most distinctly take future consequences and situations of action into account (Bratman 1984).

[388] In an early article in which decision-theoretic and AI approaches to rational agents are being combined, Bratman and others take resource boundedness explicitly into account. In that context, the subjective expected utility is taken to be 'a function of the agent's beliefs and desires.' Later, however, the cost of the process of deliberation itself is also mentioned, which seems relatively independent from the agent's beliefs and desires (Bratman, Israel et al. 1988).

[389] Bratman is aware of the risk involved in the recommendation not to constantly reconsider one's intentional structure, for he suggests that it may turn out to be necessary to formulate a historical condition which explicitly excludes the influence of extreme cases of manipulation, brainwashing and the like from it (Bratman 2003).

[390] In his project of 'creature construction' to which we referred earlier, at a relative early stage Bratman observes that there are 'substantial pressures for mechanisms' that "support coordination, intrapersonal and interpersonal, in ways compatible with these limits" (Bratman 2006b 53). Only later do capacities for deliberation about these come into play, which yield additional benefits but still need not always to be employed.

reconsideration, intentional action can normally rely upon dispositions that are the result of our having established a ramified structure of intentions.[391] Such dispositions, then, comply with the constraints mentioned earlier and are the result of the agent's long-term fostering these constraints, even though they are no longer dependent upon the reasoning, deliberating and conscious consideration that were at some point associated with them.[392] An agent's established structure of intentions and the dispositions that can emerge from it together does contribute to his constrained or sculpted space of actions, in our words, and thus facilitates the agent's cost-effective, fast, flexible and intentional action in most cases.[393]

## 3.2  Proximal intentions and cognitive mechanisms that determine anchored actions

If an expert singer has developed a distal intention to perform Saint François not in a solemn but in a nearly heroic manner though not too juvenile, he needs to recognize when is the right situation to sing a particular phrase in a way that fits that intention. Perhaps not all parts of the huge score are particularly apt for his peculiar interpretation, for instance as his vocal interactions with birds do not offer an appropriate situation to display heroism. Such an agent may have several distal intentions, all lingering simultaneously in the background and waiting for the appropriate occasion to be realized, while other must be blocked given the conflicting current stage direction. Proximal intentions are needed to navigate in a particular situation, helping him to anchor one or more distal intentions in that situation. These proximal intentions are subject to constraints of consistency and means-ends coherence, which make

---

[391] Indeed, following up on Bratman's theory of planning agency, Pollard argues for a more prominent role of habits – now not understood as mere reflex-like habitual actions. According to his account, habits are not so much dependent upon beliefs, desires, intentions and reasons but contribute to actions in a rather more embodied and embedded way (Pollard 2006). Elsewhere, such identification of an agent's ability to act with his having appropriate dispositions – as a result of having acquired habits and skills – is labelled 'New Dispositionalism' (Di Nucci 2011).

[392] Kolodny argues that it is not uncommon to praise an agent who has certain unconscious dispositions in the absence of conscious deliberation of reasons for an action. These dispositions may make 'believers and intenders' "sensitive to these reasons, either via their beliefs about them, or via unconscious mechanisms" (Kolodny 2008 390). When reasons offer constraints on actions, it is plausible that they allow processing by unconscious mechanisms.

[393] In his contribution to a symposium on consciousness, Bratman argued that a specific intention needs not to be engaged in conscious thinking continuously. He referred to the Freudian idea of unconscious intentions that still have specific content, like 'sleep with your mother'. Consciousness would than be a higher-order or relational phenomenon, that may or may not be at stake with regard to a specific intention at a specific moment (Bratman 2006a). Apparently Bratman agrees that intentional contents may contribute to an agent's actions implicitly. Such an implicit intention may still become explicitly aware to the agent. Similarly, it may have originally been consciously made. This argument also reflects his interest in developing an account of agency that is not just theoretically sound but also empirically plausible.

them different from motor intentions, which could be simply triggered by perceived affordances. Given an agent's commitment to more comprehensive action plans, proximal intentions therefore play a crucial role with their function of anchoring and specifying action. The intentional cascade framework allows us to discuss how representations of an intention to act at several levels of specificity and involving different elements are involved in the complex cognitive processes that determine an action. In the next sections we will consider whether cognitive (neuro-)scientific evidence concurs with the description of proximal intentions that we offered above. Before introducing some preliminary reflections about this, the next paragraph will remind the reader of some important lessons from our earlier discussion of empirical evidence regarding motor intentions.

Two lessons had great relevance and concerned the phases in skill learning and the modifying structure of the representations involved in this. We scrutinized empirical evidence with regard to skill learning by focusing on research of motor representations, which are capable of implicitly integrating both environmental information and options for motor action and enabling intentional action on a relatively short-term scale (Jeannerod 1997). This research provided as a first lesson that skill learning or growth of expertise corresponds with two phases that affect motor intentions: an early phase characterized by increasing efficiency of neural activations when skilled behavior is performed and a second phase, in which this behavior is strongly associated with additional activations in other neural areas, facilitating access to other representations or programs related to the skill (Petersen, van Mier et al. 1998). Second, the formation of these kludges during the process of developing a skill or form of expertise is associated with the development of chunks and templates that allow faster and more flexible processing of increasingly complex representations (Guida, Gobet et al. 2012). When we now turn to considering empirical evidence bearing relevance for proximal intentions, we may ask whether we will observe similar varieties in processing types and in forms of representations. Given the fact that proximal intentions are not restricted to processes that are by definition very fast and that escape consciousness, these varieties may present themselves in a different way. Indeed, instead of the distinction between two phases of skill learning, we will below be dealing again with a dual-process theory – similar to the dual-process theories that we've discussed in chapter II.3 ff. This turn to a dual-process theory is not made by Pacherie, who made a different choice in the present context. So why did we choose otherwise?

Pacherie does refer to some evidence with regard to proximal intentions. Accounting for the complexity of the plans involved in these, she mentions the intentional schema theory, which is built largely upon developmental, ethological and

psychological evidence. Acknowledging that this theory aims to explain intentional social interactions, her main interest is in the way that intentional schemas integrate perceptual and action related information while being hierarchically organized (Barresi and Moore 1996). Pacherie furthermore has added to her intentional cascade also forward and inverse models of motor actions determined by the intentions that figure at three different levels of the cascade. These models are used for the prediction, control and correction of actions, by being involved in cognitive processes which amount to their comparison. Like the intentional schemas we just referred to, these motor action models integrate information representing the situation with information that represents the action itself (Pacherie 2008).[394] Given that Pacherie's own account remains relatively abstract with regard to the information involved in proximal intentions and their implementation – partly because her interest is mainly in the phenomenology and experience of action and less in its determination – and given the fact that the intentional schema theory has also a relatively small empirical basis, we will below discuss whether cognitive neuroscientific insights allow us to be more specific about proximal intentions. Doing so, we will focus on a particular cognitive neuroscientific account, that offers the ingredients that are required once we intend to explain the complex functional properties of proximal intentions.

With proximal intentions, the explanatory task will be more complex than for motor intentions. Given the dual function of proximal intentions to both anchor and specify an intention in a given situation but also to block its performance in exceptional cases (Bratman 1987), their explanation must accordingly cover such divergent properties.[395] Moreover, we also noted that a person's identity and the hierarchical structure of his intentions and the constraints these put upon his intentional actions is at stake in this context (Bratman 2006b). Facilitating our explanation, however, will be the fact that for our explanation of these functional properties we will also employ the preceding explanation of motor intentions, as these are indeed generatively entreched and being employed in proximal intentions, as well. Integrating elements of the explanation of motor intentions, we might expect the explanation of proximal intentions to be more complex both in terms of processing and of representations. Let us first take a quick

---

[394] Elsewhere Pacherie adds to her model a distinction between two different forms of control – following an analysis offered in (Buekens, Vanmechelen et al. 2001) – namely: tracking control and collateral control. Tracking control is engaged for flexibly adjusting the motor action in order to succesfully reach the goal, collateral control is controlling for undesirable side effects (Pacherie 2006). Note that the latter form of control could amount to blocking the execution of the action, which is one of the functions of proximal intentions according to Bratman (Bratman 1987).
[395] Concurring with Bratman's intention-action principle which holds that an agent controls his actions via his intentions (Bratman 1987), many empirical studies of the control and determination of action are comparably interested in the representations and reasoning that are involved in the agent's intentions.

look at some notions that have been proposed to account for some properties we associate with proximal intentions.

During the last century and particularly its last half, researchers have developed constructs that enabled them to explain the quite divergent functional properties of intentional actions, the changes that occur to these due to development and learning, and the errors or deviations that do obtain in exceptional cases. Generally speaking, these constructs were primarily presented in terms of the representations and cognitive strategies involved, with their neural implementations being explained only more recently. An early example is the 'schema' construct. It was put central by Bartlett in 1932 – who actually preferred the term 'active developing patterns' to the word 'schema' - explaining how an individual always integrates novel information into representations that have developed from past experiences, instead of his merely assembling separate memories (Bartlett 1995 [1932]).[396] Such hierarchical structure emerged not only from empirical but also from computational studies. Indeed, based as they were upon not only psychological but also on computational and simulation studies, the term 'scripts' was suggested for those higher level representations of temporally extended actions (Schank 1980).

These knowledge structures have been not only applied to cognitive processes but also to other domains, like Piaget and others did when they used schemas to account for developments in behavior, language and thought (Arbib 2003b). Indeed, the study of skill learning did equally reveal the importance of hierarchically structured representations, with 'plans' becoming increasingly stable and governing over time complex habitual actions in an automatized fashion (Miller, Galanter et al. 1960) Moreover, this employment of hierarchically structured schemas – implicitly in most cases – could not only explain the structure of both behavior and language, but also explain exceptional cases or disorders in which the serial order of actions is affected, for example (Lashley 1951).

However, research did not only demonstrate how the representations involved in various forms of intentional action are complex and hierarchically structured, it did also suggest that more than just a single cognitive process might be involved. Indeed, such representations are considered as frameworks which are the result of the integration of several sources of information and which are also employed by different

---

[396] Interestingly, Bartlett explicitly contended that schemata also include social and cultural information. This social aspect, though, has been left out of the schema concept when it was taken up in cognitive science some decades later only to reenter the research more recently (McVee, Dunsmore et al. 2005). Indeed, in our analysis of distal intentions below, which refer to comprehensive social and cultural influences will be present as well.

processes simultaneously (Minsky 1975). An example of such parallel employment is when schemas are not only being used in the initiation and determination of actions but also play a role in the ongoing control and correction of these (Arbib 1981). Or, as another computational model accounting for psychological evidence suggests, information is integrated that has been retrieved from both long-term and working memory, which underlies both the flexibility and increasing stability of cognitive, language and behavioral skills (Anderson 1983).

What can be gathered from this short glimpse of some relevant theories is that explanations for the complex functional properties of expert behavior are indeed striving to combine insights regarding both representations with their complex structures and different processes with potential interactions. We have chosen to organize our discussion of a plausible explanation of our account of an agent's sculpted space of action by focusing on an influential theory and corresponding model of action determination that has been built upon the notions that we just mentioned, has been tested against empirical findings, and has accordingly been subject to proposed modifications. Norman and Shallice have developed a kind of dual-process theory of 'willed and automatic control of behavior', which has been primarily developed as a computational model for every day actions and for action errors or disorganization that can be seen in healthy subjects and patients. Based upon clinical and other observations, the theory meant to "account for the ability of some action sequences to run themselves off automatically, without conscious control or attentional resources, yet to be modulated by deliberate conscious control when necessary" (Norman and Shallice 1986 378).[397]

Taking up several features of the constructs that were mentioned just before, actions are in this theory, too, taken to be hierarchically structured, with lower level motor schemas[398] containing sensory-motor mappings that determine muscular movements and higher level scripts determining the ordered and adequate performance of such motor schemas when performing a complex and temporally extended action (Cooper

---

[397] In our discussion we will not just rely on the original formulation in (Norman and Shallice 1986) but refer to the updated and expanded version of the theory as for example presented in (Cooper and Shallice 2000 ; Cooper and Shallice 2006).

[398] The notion of motor schema is similar to that of Jeannerod's motor representation, referred to in our sections on motor intentions (Jeannerod 1994): it integrates both information about affordances and about specific motor actions. In the current context, however, the additional question is how several such motor schemas are carried out in an appropriate order such that the performance of a complex action is enabled by them (Cooper and Shallice 2000).

[399] Indeed, the updated and implemented version of Norman & Shallice's model does incorporate the hierarchical analysis of everyday action and disorganization errors as presented in (Schwartz, Montgomery et al. 1995 ; Schwartz, Reed et al. 1991). As a result, the updated model distinguishes five levels of action control, ranging from, for example, the upper level of 'morning routine' to the lowest level of 'take cream container', belonging to the sub-routine of making coffee as a part of breakfast (Cooper and Shallice 2000).

and Shallice 2000).[399]Such representations are the result of learning and experience and include information about relevant environmental trigger conditions like objects that can be manipulated, about goals or end states to be reached, and about the relevant motor behaviors that might realize such end states. The result of the complexity of these representations is that an action representation can be activated via a variety of triggers, as our expert singer's representation for singing might be activated via the perception of a mandoline or of Zerlina, for example.

Specific for this theory is the assumption that such action representations are established with learning and experience, and then figure in an 'interactive activation network'. The latter implies that action representations can be considered as nodes in a complex network, with each node having a variable activation level that is determined by many different factors, including environmental triggers, motor effectors, and other ongoing processes. Moreover, each action representation can also influence the activation levels of other representations, for example by increasing the activation level of its own lower-level, component actions representations and – conversely - by inhibiting the activation of representations of competing actions. A routine action is carried out once a particular representation is activated beyond a given treshold and has outcompeted rival ones that have become activated as well.[388] As a result, well-learned actions can accordingly be performed without any influence of conscious or deliberate control, simply as a result of a change in a representation's activation level (Shallice 1988).

Apart from the 'contention scheduling' (CS) that is responsible for automatic control of actions, this model describes a 'supervisory attentional system' (SAS) that is responsible for deliberate action control. According to the model, routine action and cognition would not require such control, leaving top-down control necessary only "if error correction and planning have to be performed, if the situation is novel, or temptation must be overcome" (Norman and Shallice 1986 382). in novel tasks or in complex tasks dorsolateral prefrontal cortex activation is increased. This is interpreted as activation associated with the acquisition or generation of rules that help to select a response (Crescentini, Seyed-Allaei et al. 2011 ; Frith, Friston et al. 1991).[389] In its original formulation, this SAS exerts its effects by a top-down process that amounts to focusing attention on a particular action representation or relevant features of it.

---

[388] This process of the selection of an action representation under specific conditions is described more in detail as 'contention scheduling' by the authors and consists mainly of activation and inhibition of potentially performed representations (Norman and Shallice 1986). We will only present this process to the extent that is necessary here.

[389] Indeed, motivated partly by this CS/SAS model, a study investigated subjects' employment of rules that constrain the search space in a response task. In that context, the authors use the phrase 'sculpting the response space', which has inspired the title of this dissertation (Fletcher, Shallice et al. 2000 ; Frith 2000).

In doing so, it primes or biases the corresponding action representation and thus can modulate the CS process. Further elaborations have brought about a subdivision of SAS processes, all of which are still considered to modulate the activation levels of action representations and in that sense cohere as a supervisory system.[390] Different supervisory subprocesses that are dinstuighed are the spontaneous generation of an action representation, the specification of an action in order to solve a particular probem or the retrieval from memory of a particular action representation (Shallice and Burgess 1996). As these processes cannot directly influence the performance of an action but only indirectly by increasing or decreasing activation levels of action representations, top-down influences on an agent's actions are always competing with the other processes that determine his eventual performances.

With these two processes of 'contention scheduling' and 'supervisory attentional systems' in place, let us consider whether they can provide some more insights in the properties of proximal intentions that philosophical analyses have presented to us. Moreover, they may perhaps add to these analyses, while we may, conversely, make suggestions for further developments of the cognitive neuroscientific theory and model.

### 3.2.1  Processes for the resolution of conflicts between action options

In section III.3. we reminded the reader of the possibility that a single opera scene might offer many affordances for action, which suggests that more than just a single motor intention might be activated. How would our expert singer be capable of fast and adequately selecting only one motor intention for performance? Even more problematic seemed the situation in which he would have to block the performance of a habitual motor intention in an exceptional situation, as when a stage director requires a senior Don Giovanni or a heroic Saint François. Finally, we asked how coherence over time between intentions was created, given that motor intentions did not appear to provide for this.

We learnt from our subsequent discussion of arguments by Frankfurt and Bratman that selecting an option for action from the many options available in any given situation is a complex task which asks for processes and choices that far transgress the

---

[390] In singing at least four processes have been distinguished on the basis of collected evidence: auditory perception, a decision making process for retrieving and selecting an appropriate representation, the execution of the corresponding motor plan and finally the evaluation of the outcome of it – the sung tone or melody (Hutchins and Peretz 2011). Experiments with expert singers show their representations to be hierarchically structured, depending on their expertise with musical structure and performance (Zurbriggen, Fontenot et al. 2006). The complexity of these representations can be more complex than those that have been discussed in the context of motor intentions, which are limited in temporal extension and play their role mostly implicitly.

boundaries of any particular situation. Solving this task and resolving the potential conflict between action options requires an agent to organize and coordinate his intentions for action such that a coherent structure determining these intentions emerges. This coherent structure consists of 'constraints' that an agent takes upon himself and that limit the space of options that are open for him to perform (Frankfurt 1999). Further elaborating the coherence of this structure and analyzing how it affects an agent's actions in various situations, Bratman has noted a tension between two requirements to which an agent has to answer and which befall to proximal intentions to fulfill. On the one hand, an agent should refrain from costly and counterproductive reconsideration of the coherent structure of his comprehensive action plans, while also being prepared to block the performance of an action that belongs to such a plan, if there is a higher-level or distal intention that demands such a deviant response in a given – emergency, for example - situation (Bratman 1992a).

The dual-process CS/SAS model initiated by (Norman and Shallice 1986) provides several ways in which an agent can comply with the requirements that the analyses have presented – implicitly and explicitly. Did a motor intention integrate specific environmental information and a motor response option, the action representations involved in the current model can contain more information. With that, they are potential implementations of the proximal intentions discussed above. As a result of learning and experience, the scripts and schemas included in this model figure in comprehensive hierarchies of actions and also determine sequences of actions. Given the 'interactive activation' that determines the connections within these hierarchies and which is to a large extent the result of experience, there are continuously many vertical as well as horizontal interactions that contribute to an agent's sculpted space of actions. Vertical interactions obtain when a comprehensive action script has been initiated, for example by singing a Don Giovanni performance. In that case, the activation levels of a set of component action representations at multiple hierarchical levels is being increased. Horizontal interactions also occur, as when the activation level of competing representations is being inhibited or decreased at the same time (Cooper and Shallice 2000).

With such hierarchical activation patterns established over time, our expert singer will usually be able to perform coherently once the appropriate comprehensive action representation has been activated. One consequence of this is that motor intentions that do not belong to this representation will have less chances of becoming activated: in a given situation both irrelevant environmental triggers and inappropriate motor representations are normally less activated or sometimes even inhibited (Cooper and Shallice 2000).[391] When Don Giovanni is trying to seduce Zerlina, for example, it is

most unlikely that he will use the scolding voice with which he will later answer Donna Elvira. Even more unlikely is that he will meditate upon a cross when having his dinner near the Commendatore's grave as our singer might do when playing Saint François or that he will accidentally pick a handbow from the wall as if he was singing Guillaume Tell. Indeed, with the addition of extra pre- and postconditions to the hierarchically organized component actions, the 'contention scheduling' model has received extra features that support his appropriate performance. According to these conditions, an action (component) will be activated by particular environmental features or by the completion of a preceding component. Conversely, the action will be inhibited if a particular goal or environmental situation obtains, even when the agent has not himself contributed to this to happen (Cooper 2002).

The study of deviant behaviors in healthy agents and in patients has influenced the development of this CS/SAS model and has also contributed to insights regarding the coherence and the flexibility of normal actions (Schwartz 2006 ; Schwartz, Montgomery et al. 1995). Results show that there are many ways in which actions can become disorganized, ommissions of action components occur, agents suffer from perseverations or unnecessary, repetitions of actions, misuse of objects for specific actions, and so on. These errors affect the 'intermediate level of organisation', as they happen in a time frame that would allow conscious control (Cooper and Shallice 2000), which distinguishes our proximal intentions from motor intentions. The current model offers several explanations for incoherencies visible in these errors. Since the activation level of a particular action representation is based upon the summation of activation levels that depends from excitation or inhibition through neighbouring action representations, from bottom-up excitation by environmental triggers and from top-down excitation via an activated higher level action representation, there are multiple ways in which this coherence and flexibility can be negatively affected. This is the case when the 'parameters' of interactions that are prevalent in a patient's CS mechanisms are set at a level that is so high or low that inappropriate excitation or inhibition obtains (Rumiati, Zanini et al. 2001). Obviously, the other side of the coin is that an agent's coherent acting can equally be explained via this model, even though it may be in terms different from the ranking or prioritizing of intentions for action that we discussed in section III.3.1.1.

---

[391] The theory about 'Structured Event Complexes' is developed to some extent in discussion with the CS/SAS model, although it has distinct ideas about neural implementations of the processes it refers to. Similar to both approaches, though, is the idea that interactive activations between component representations determine the selection of a hierarchically structured event during cognition or action (Grafman 1995). The SEC theory is also developed to account for the observation in healthy and normal agents that some SECs will be more rigid in their structure, whereas others are more flexible and allow ample adaptation to environmental conditions (Grafman 2006).

Most important, the compositions of the action representations' hierarchies and all activation levels involved are largely dependent upon the individual agent's development, learning and experience.[392] Given the Hebbian learning that is involved in these processes, all compositions and activation levels are obtained over longer periods of time.[393] The dynamics involved in these allow for many different influences, ranging from exposure to particular environmental triggers to deliberate choices that favor some over other actions and also provide quite some flexibility to each inidividual (Cooper and Shallice 2006).[394] For example, action observations and modelling studies suggest that the association between environmental triggers – comparable to affordances - and particular actions is a result of individual exploration and experience and allows adaptation to novel environments (Cooper and Glasspool 2001).[395] However, the question is whether the model allows the kinds of interaction between 'automatic and willed control of behavior' (Norman and Shallice 1986) necessary for prioritizing to occur: how can an agent deliberately control the automatic behavior that results from the CS mechanisms?

Obviously, where an individual agent will gather experiences and thus develop his interactive activation networks also in the absence of deliberate control, he has also the capability of supervisory control that allows him to exert some influence on these networks. In the initial formulation of the CS/SAS, it is the agent's 'attention to action'

---

[392] This is not to deny that there are many representations or components thereof that will be shared among individuals. Remember that in the previous Part we referred to a neuro-constructive account of development. Such an account emphasizes the brains 'embodiment' and 'ensocialment', which contributes to shared representations between individuals irrespective of their individual learning trajectories (Mareschal, Johnson et al. 2007). Before that, Karmiloff-Smith already emphasized the importance of innate prespecifications that facilitate the newborn's learning, challenging as she does both "Fodor's anti-constructivist nativism and Piaget's anti-nativist constructivism" (Karmiloff-Smith 1994 693). Apart from a shared social environment and stimuli that stem from this, these innate prespecifications also contribute to partly shared representations among agents from an early stage on.

[393] For example, switching to another task or course of action requires an agent to use his intentions for reconfiguring the activated action representations, which does require extra cognitive capabilities and resources. This explains why often agents persist in completing an initiated action (Goschke 2000). Task switching can be distinguished along a gradient which varies from exogenous control that is stimulus driven to endogenous control and has been associated with a posterior-anterior gradient along the prefrontal cortex (Kim, Johnson et al. 2011). As the CS/SAS model allows that the initiation of a comprehensive action activates simultaneously alternative component actions with their own trigger and goal conditions, expertise can accordingly indeed allow such task reconfigurations. However, activation through trigger conditions is powerful and it requires strong supervisory processing to override these (Norman and Shallice 1986).

[394] Obviously, there are many constraints at work that limit this flexibility. For example, house flies have to cope with physical constraints in their behavior that favor some behavioral sequences over others (Dawkins and Dawkins 1976), which holds to humans as well. Among others, however, cultural and individual constraints  also determine the representations that drive human behavior (Grafman 1995).

[395] The template structure that we noticed to be present in motor intentions returns in the present context, too. The hierarchical action structures at stake here does not demand that representations are determined up to the lowest level, leaving room for flexibility in the motor movements to perform a certain action.

that can affect already existing preferences or help to respond to novel action situations (Norman and Shallice 1986). Later elaborations have distinguished between different supervisory processes that support control of actions by modulating activation levels such as spontaneous generation of an action representation, the specification of a goal-directed action or retrieving a particular action representation (Shallice and Burgess 1996). What he in fact is doing through such processes is priming targeted features of his own CS mechanisms.

Several strategies for controlling otherwise automatic actions have been discussed in section II.3.1.3 in the context of dual-process theories, which share commonalities with the current CS/SAS theory. A way of self-regulation would accordingly be for an agent to articulate a rather specific intention that he aims to implement in a situation in the near future, like responding to a particular environmental conditions with a specific action (Gollwitzer and Sheeran 2006). Using a converse strategy, he could prepare himself to inhibit a certain action by engaging in counterfactual thought which primes him for action options that are alternative to his routine ones by modulating his attention to otherwise unnoticed environmental triggers or his evaluation of undesirable action consequences (Galinsky and Moskowitz 2000). Indeed, part of the supervisory processes of the CS/SAS model are also processes that monitor and evaluate actions (Shallice and Burgess 1996).

Even though these supervisory or attentional processes are cognitively demanding and recruit resources that automatic action does not require, agents are capable of influencing their actions by employing them. Yet for effectuating lasting changes in his action routines, it would require him to realize long-term intentions to modulate these. On the other hand, by the time an agent has developed over time the action representation hierarchies and activation levels that conform to his intentions, 'contention scheduling' does allow him to act according to his intentions even in the absence of conscious control. This raises the question how blocking a routine action can take place, it being one of the roles that proximate intentions play.

### 3.2.2  Blocking a habitual action according to the CS/SAS model

As proximate intentions should among other functions allow an agent to block the application of a distal intention in a specific situation we should hope that hierarchical action representations do not rigidly require their comprehensive execution upon an agent from the moment of its initiation. In terms of Bratman's own example of not buckling up in a case of emergency (Bratman 1987), we would hope that the comprehensive script for driving, including as a component action one's buckling up, allows to be modified. Driving, for example, would then be subordinated to the

superordinate action of bringing someone as quickly as possible to the hospital, which corresponds with adjustments like the following: component actions that are irrelevant for the superordinate action can be left out completely; objects that distract from the primary task should receive less attention; the physical well-being of the patient should receive some extra attention. As a result, the activation levels pertinent to component action representations, to environmental triggers, to action goals – or criteria for their fulfillment – require modulation. Such modulations can be hard to reach, since it may be difficult for an experienced driver to omit buckling up before driving. A comparable flexibility for our expert singer would make us expect him to be flexible in repeating a passage of his score, to interpret it differently or to sing his aubade not under a balcony but on a boat instead. In contrast, we may expect him to have great difficulty in changing particular notes in a fast melody at will. Apparently, we do assume that some action components are more stable and less flexible, then others. The study of errors or disorders that are visible in action performances can learn us more about our assumption or expectation – indeed, most errors in patients are considered to be extreme forms of normal errors that are made by healthy agents (Cooper 2002). What can we learn from errors and how do we normally avoid errors or unintended actions to occur?

Similar to the relative flexibility of motor representations that we have found in section III.2.2.1, relying on a template structure with open slots that permit variability, we should expect the representations of proximal intentions to allow some flexibility. Indeed, in a theory of action schemas preceding our CS/SAS theory the difference between closed and open skills was made. Whereas closed skills were taken to rely on a constant environment that permits identical performance of a particular action, open skills require a skilled performer to adapt his skill performance to environmental changes (Schmidt 1975). This and other types of variability have been integrated in Norman & Shallice's comprehensive network model of interactively activated components and are also responsible for action errors. Let us first look at the representations involved and how they would allow flexible adjustments, before focusing on the processes that potentially would carry these out.[396]

To begin with, different action hierarchies can share component representations,

---

[396] Lacking in the CS/SAS model are affective or motivational processes. In the –somewhat related - theory of Structured Event Complexes, an explanation is offered of agent's retrieval of action events from memory and how this impacts upon his behavior. It explicitly associates motivational and effective values with action components, explaining why actions usually appear to be satisyfing and why some patients would repeatedly engage in apparently aimless actions – like patients suffering from utilization or compulsive behavior (Huey, Zahn et al. 2008). The CS/SAS model could be expanded by letting such motivational processes also – indirectly, perhaps - modulate specific activation levels.

since a particular component action can contribute to more than just a single comprehensive action. Indeed, based upon this characteristic, so-called 'capture errors' can occur, when a particular action sequence shifts to another action due to the fact that they share a common action compnent (Cooper and Shallice 2000): boiled water is being poured in a tea pot instead of being used for making coffee, since both are components of the action representation of pouring boiled water. In this case, the activation levels pertaining to the irrelevant object (tea pot) and corresponding action goal (making tea) are apparently not decreased sufficiently to avoid this error from happening. For another, completing an action does at times require an alternative component that deviates from routine components to become flexibly integrated its performance (Cooper 2002): a flexible cook would not hesitate to improvise by using a tea pot when making coffee with no coffee pot present. Are such resources of variability of performing a complex action based upon the complexity of the interactive activation networks with their distinct roles for components like environmental conditions, higher-level action intentions, and lower-level motor intentions - random noise is still an important additional source.[397] Without such variability the choice between two equally activated, competing alternatives – using a left or right hand, for example – could lead to an impasse. Deciding how to specify the action would then require supervisory processing which would slow the action down and demand extra cognitive resources which are unnecessary if variability avoids such a stalemate. Random noise – or another source of variability, we could add - would ensure that activation levels of both hands are only rarely equal (Cooper and Shallice 2000).

With the representations allowing modifications via modulation of activation levels pertinent to component action representations, to environmental triggers and to action goals – or criteria for their fulfillment – the question remains which processes may be capable of reaching such results. As we've noticed above, there have been distinguished several supervisory processes that allow willed control of an action through overriding

---

[397] There is increasing evidence that random noise enhances the sensitivity and flexibility of processes in the brain. In contrast to systems characterized by linearity of their operations and interactions, so-called 'stochastic noise' can enhance the performance of the brain – operating largely as a non-linear system – as was shown in a random noise stimulation experiment with healthy subjects on a perceptual learning task, for example (Fertonani, Pirulli et al. 2011). More generally, it has been proposed that 'chaotic itinercy' occurring at several mechanism levels – from neuronal assemblies to social interactions – can explain how stability and variability appear to be interdependent (Kunihiko and Ichiro 2003). Similarly, Freeman has demonstrated how the complex dynamics of oscillations that obtain due to different neuronal firing patterns can lead to a signal to noise ratio that offers minor stimuli a limited chance to greatly affect brain processes (Freeman 2000). There are many sources of variability in the brain – indeed, the occurrence of stabilized patterns of synchronized neuronal activations should perhaps be more surprising than the occurrence of the opposite. Accordingly, it has been pointed out that most brain studies focus on task-related processes that explain some 5% of the brain's energy consumption, whereas spontaneous neuronal activity uses most of it. Fox & Raichle argue that instead of calling this variability 'noise', it to a large extent correlates with activity of the default mode network (Fox and Raichle 2007).

and modulation the process underlying routine actions. Supervisory processes can be recruited at the different phases or stages of an action performance, depending upon its temporal and situational conditions. A coarse distinction can be made between the preparation, implementation and assessment stages, each characterized by distinct processes contributing to the action (Shallice and Burgess 1996). For example, if time permits and the situation is familiar, an agent may set novel action goals or configure an appropriate alternative action representation - prepare himself for using another tool or omitting a particular component action - and thus modulate the representations involved in his action routines. Alternatively, he may prepare himself to note particularly well whether a specific component action goal has been reached, depending on which another component must not be performed (Shallice and Burgess 1996).[398]

According to the original CS/SAS model, such blocking of an action would require supervisory processes depending upon frontal lobe activity in constrast to automatic action (Norman and Shallice 1986). As a consequence, the limitations of frontal processes in terms of work load and speed would affect such willed actions, limitations that were also at stake in controlled versus automatic processing that we've discussed in the context of dual-process theories in chapter II.3 However, comparable to the discussion about the distinctions and relation between controlled and automatic processes, the relation between CS and SAS is under debate.[399] Research of action errors in in patients, for example, has demonstrated that some of their routine behavior is also affected by frontal damage, even though the original model suggested that such damage should only affect supervisory processes that take place in the frontal lobes. Apparently, it was concluded, routine actions do also involve some monitoring and correction, without necessarily requiring costly supervisory processes (Cooper 2002). A study of the effect of forming implementation intentions preceding the performance of stimulus-response actions confirmed that a supervisory process can preliminarily influence automatic action without slowing it down. A goal-intended action as a response to a novel stimulus was performed faster after forming such implementation

---

[398] From studies of the regulation of automatic behavior, it is known that priming can affect an agent's future perception of critical environmental features or his entertainment of certain goals, for example (Macrae and Johnston 1998).

[399] Research in subjects with high and low susceptibility to hypnosis suggests also that the supervisory processes in those with high susceptibility have weaker connections to CS processes. Subjects who are highly susceptible to hypnosis were more flexible in shifting their attention and responses in a Wisconsin Card Sorting Test, suggesting less influence of supervisory processes. In another test these subjects were also less capable of recalling details from the performed task (Aiktns and Ray 2001). These results confirm that supervisory processes can indeed impede flexible routine actions (Norman and Shallice 1986), but also that they may be required for establishing explicit representations.

intentions, even in frontal patients that suffer from utilization behavior or other disorders that tend to negatively affect response behavior (Lengfelder and Gollwitzer 2001). These results are taken to imply that automatic processing can be modulated at once following the reconfiguration of an intended action preceding a task.

In sum, studies of the interaction between automatic and controlled processing do indeed confirm that agents have the capability to block or modulate a proximate intention. However, they do also point out that multiple options for such modulation are available, each depending upon different processes and each targeting different components of the action representation and its underlying network. Corresponding to this, there are multiple constraints that limit the flexibility which is required from a responsible agent as he emerges from the philosophical analysis. This analysis has itself focused on some specific constraints of proximal intentions. Let us consider these as well.

### 3.2.3  Multiple processes and the constraints for anchoring an intention in a situation

In section III.3.1.3 we discussed constraints for anchoring an intention in a particular situation, which requires assessing whether that situation is appropriate for its execution and then specifying the action adequately. These constraints amounted to a demand for the action's internal consistency and for its specification in terms of "subplans concerning means, preliminary steps, and relatively specific course of action, subplans at least as extensive as I believe are now required to do what I plan" (Bratman 1987 31). Given the different processes involved in action control according to the CS/SAS model, the implementation of these constraints is manifold.

Concurring with the latter constraint, the model does ascribe to action representations a hierarchical structure which integrates many elements, among which environmental triggers, objects, component action representations, pre- and post-conditions, as well (Cooper and Shallice 2006). Once an action has been often practised, this structure and its elements are connected via interactive activation levels with each other. Depending upon the expert's experiences, his action routines will depend upon an elaborate representation with a great number of elements that can potentially be activated, allowing him to anchor the activated action representation rather flexibly in any given situation. A crucial feature of the model is that it also includes optional elements at several levels of specificity and of various kinds. For example, if the temporal order of an action sequence is flexible, the lateral activation and inhibition levels of neighbouring component action representations will be less decisive and thus allow the sequence to be influenced by the perception of relevant

objects or the occupation of his hands, which can consequently – bottom-up - influence these activation levels (Cooper, Schwartz et al. 2005). When Don Giovanni and Leporello change clothes, the order of their dressing is dependent upon the availability of a cape or trousers, upon their having free hands or balancing upon a single foot, and so on: the anchoring and specification of the action is facilitated by the fact that the contention scheduling process can dynamically operate within a space of actions characterised by a complex structure of interactive activation levels that is sculpted through the accumulation of experience.

This explanation of how a proximal intention is carried out is quite different from the philosophical rendering of it. We've discussed how the philosophical account emphasizes that we should expect proximal intentions to operate between the domains of – implicit, experience based – motor intentions and distal intentions. Proximal intentions are then responsible for the process of specifying a distal intention such that it can be anchored in a given situation or, conversely, blocked in an exceptional situation. Without proximal intentions fulfilling these roles, it was argued, it is difficult to conceive how intentional actions can be performed. The implementation of these roles in terms of representations and cognitive and neural processes has shown that there is not a neat mapping of levels of intentions to the component processes suggested by neuroscientific research. For each intention type depends upon an explanatory mechanism of great complexity, containing distinguishable components which have not always recognizable counterparts in the results of philosophical analysis. Indeed, this lack of comprehensive correspondence between the empirical and philosophical accounts have to do not only with differences in aim and goal, but also with differences in the relevance attached to results of the other account.

Notwithstanding these differences, the CS/SAS theory and model can account for observations and simulations of routine actions that are in agreement with functional properties as presented in the philosophical analyses. A more direct similarity between the two perspectives is when explicit action representations are used by supervisory processes to control an action. Even though these processes are only capable of modulating routine actions, according to the model, this modulation can indeed lead to blocking an action wholesale, or altering targeted elements of it. Obviously, for such targeted modulations, the agent must indeed be experienced as he must be able to configure an alternative course of action, using a different object, following another sequence, and so on. A dissociation between such distinct supervisory processes has been proposed, together with different neural implementations for them (Shallice and Burgess 1996). Even though research has shown that imagery can also support such action preparations or modifications (Kosslyn 2008)[400], it is assumed that verbal

representations will usually play an important role.

With regard to the former constraint, referring to the consistency of action, this would also be supported by the strength of the interactive activation patterns of an agent. Indeed, action disorganization or utilization behavior has been simulated by taking connections or elements of the comprehensive representation out or by adding random noise to the model, which amounts to blurring patterns of associated activation levels. As a result, the lateral inhibition of two inconsistent component action representations is decreased, for example, making it more likely that both of them are being performed.[401] These simulations have been compared with problems in neural connectivity – deviant neurotransmitter or receptivity levels, for example - and with the presence of specific neural lesions, responsible for the absence of elements in the relevant representations (Cooper and Shallice 2000).

With the later addition of pre- and post-conditions to the representation of each component action, the processes enhancing an action's consistency were further expanded. As these conditions can be considered to add to an action representation's activation level by taking into account activations based upon the presence of triggering (pre-)conditions and of inhibitory (post-)conditions, they further constrain its activation (Cooper 2002).[402] Given that these triggering situations and goal states are also relevant for maintaining consistency in routine action, the originally sharp distinction between supervised and automatic action processes has been weakened.

Indeed, research of frontal lobe patients demonstrated that different types of errors in routine behavior occur in correspondence to different lesioned areas, thus suggesting that routine action does also rely on some type of supervisory processing (Schwartz, Montgomery et al. 1998).[403] This finding did concur with another study

---

[400] Kosslyn explicitly refers to mental imagery as a way to 'cognitively restructure' a stimulus or event, including its affective properties (Kosslyn 2008).

[401] Utilization behavior has indeed been associated with a lack of inhibition of impulsive action due to lesion of the frontal lobes (Archibald, Mateer et al. 2001). Indeed, the author who coined the term 'utilization behavior' emphasizes how subjects with this disorder are 'abnormally dependent' upon stimuli from the social and physical environment (Lhermitte, Pillon et al. 1986). The lack of inhibition and increased sensitivity to external stimuli together constitute the disorder's most important symptoms.

[402] Presumably, the perception and recognition of relevant environmental triggers and the recognition of having reached an action's goal state depend upon activations of distinct neural networks. Mirror neurons can play different roles in these processes. Especially for object related actions, mirror neuron networks have been found that are sequentially activated during different phases of such actions – from the trigger phase of an action to its goal state. For that reason, some mirror neurons are said to be 'logically related' (Iacoboni, Molnar-Szakacs et al. 2005).

[403] Denying the sharp distinction between two processes responsible for automatic and willed control of action, Schwartz et al. explain the action errors in automatic control via a lack of cognitive resources (Schwartz, Montgomery et al. 1998). Such a lack perhaps obtains more often in automatic control, as distractor objects may play a greater role under such conditions than in willed control, for example.

of differentially impeded action representation capabilities in patients. Errors in the organization of an action script were found to be associated with frontal lobe lesions, whereas the generation of a new action script appears to rely on semantic networks that are located in more posterior areas (Sirigu, Zalla et al. 1995). Apparently, in both types of processing, action consistency is supported by several cognitive processes that each contribute to different forms of action consistency like its appropriate initiation, its completeness, its correct order and its adequate completion.[404]

In sum, the adequate activation of a comprehensive, hierarchical representation of a complex action together with the appropriate activation levels that connect action components, triggers and goal states with each other, can be impeded in multiple ways as it is dependent upon a rather large network of neural areas interacting with each other. Flaws in these processes are observable in an agent's behavior, for example in his failure to perform a sufficiently specified action or to do so in a consistent way. As noted before, there are strategies an agent can employ to prevent this, like when he forms implementation intentions (Gollwitzer and Sheeran 2006), engages in counterfactual thinking (Galinsky and Moskowitz 2000) or employs mental imagery (Kosslyn 2008). However, such strategies require that an agent is capable of a preliminary specification of relevant action elements. For only on that basis is it possible for an agent to modify or reconfigure such an action. This reminds us of our extensive argument in Part II, that although expertise corresponds with kludge formation and some modification of the representations in use, it cannot imply the loss of the capability to articulate the representation that is used in the expert's skill. Without the latter, an expert would be less capable of modifying, adjusting or correcting his actions than a novice, which appears to contradict the notion of expertise itself.

Let us consider in the final section devoted to proximal intentions some neural evidence pertaining to them and see whether we find in this context kludge formation to play a role as well, affecting the mechanisms responsible for them.

### 3.2.4 Proximal intentions and some evidence concerning their neural implementation

The motor representations and the neural processes that are involved in motor

---

[404] Research in comprehension of texts during reading has motivated the development of an interesting 'landscape' model, in which interactive activations – here of concepts - play an important role as well. The authors demonstrate both in empirical and simulation studies that readers employ two mechanisms that together constrain incorrect or incoherent interpretation: associated concepts are being activated while a parallel coherence-based retrieval process can help to detect errors (Tzeng, Broek et al. 2005 ; van den Broek and Kendeou 2008). Given that processing of language and action overlap to a large extent with each other, it would have been interesting to further compare this 'landscape' model with our model of a 'sculpted space'.

intentions were found to undergo changes as a result of gaining expertise, which leads to what we've been calling 'kludge formation'. Earlier in this chapter we've described how a phase of increasing efficiency in neural activations is followed by a phase in which representations become increasingly associated with other representations or processes that allow a wider range of exploitation of the motor representations. The representations itself do develop into more complex ones, as well, and develop a template structure in which information is compressed in the form of chunks in combination with open slots to allow for the integration of variable information. With proximal intentions and their involvement in anchoring and specifying an action in a given situation and in blocking it in exceptional cases and two different computational processes responsible for them, it is plausible to expect that their neural implementation is more complex, too.

Norman and Shallice's theory of willed and automatic control of action has also been discussed with regard to its plausible neural implementation. Obviously, Hebbian learning and other processes that affect connectivity are put central, since the theory posits that every action depends upon a specific configuration of activation and inhibition levels of a hierarchically structured set of action representations, eventually leading to the selection and execution of a particular action (Norman and Shallice 1986).[405] In the theory's expanded version we learn that after a period of skill learning a hierarchical action representation has been established and that this complex action is also assumed to be somehow represented neurally, for "even in a domain as loose as the organization of everyday routine action, one cannot simply dispense with *units* or *discrete states* representing action subroutines and goals (Cooper and Shallice 2006 906, italics added). However, the specification of this neural implementation of an action representation is here left open.[406] At the same time, the theory also predicts the emergence of such representing units or states while equipping these representations with various kinds of open slots. For it is with these open slots that an expert action – characterized by stable and efficient activation patterns -  can flexibly vary in the sequence of its component actions or can respond to several potential triggers or make

---

[405] Indeed, it is suggested that when the CS/SAS model is used for simulation studies: "a parameter may be taken to correlate with the level of a neurotransmitter", or  a "parameter may be related to the ratio of the connectivity from the two activation pathways to schema nodes" (Cooper and Shallice 2000 323).

[406] This remark of the necessity of the representation of action subroutines and goals by units or discrete states is aimed against an alternative computational model that aims to avoid a representation of a action's hierarchical structure. It is implemented in a recurrent connectionist network to simulate action sequences and disorders of action, making it a plausible alterantive according to (Botvinick and Plaut 2004). It is debatable, however, whether this alternative model can equally account for the supervisory processes or other strategies for action modulation that rely on explicit and specified action representations (Cooper and Shallice 2006 906). We've discussed a similar debate with regard to learning and development and the role of – both implicit and explicit - representations in Part II.

use of different effectors (Cooper and Shallice 2006). As a result from this first take on their neural implementation, it appears that proximal intentions – mediating between motor and distal intentions – are also characterized by the formation of kludges and the development of representations with a template structure. Let us consider more specifically how these requirements could be neurally implemented.

Neural implementation of the original theory was thought to require two distinct neural mechanisms, one for automatic action control and another for the supervisory processes involved in willed control (Norman and Shallice 1986). This distinction was challenged by studies which showed that frontal patients were not only impeded in such willed control but also in their routine actions, suggesting that the two mechanisms were less distinct than previously assumed (Schwartz, Montgomery et al. 1998). Such pathological evidence in combination with animal, developmental, and neuroscientific evidence has led to the proposal that prefrontal cortical activations are generally involved in developing, storing and processing hierarchically structured representations that are also being employed during different types of action processing. These representations are called 'structured event complexes' or SECs and they can vary in several dimensions (Grafman, Sirigu et al. 1993).[407] As this SEC theory challenges the CS/SAS theory particularly with regard to its neural implementation and not so much with regard to its assumptions regarding the representations involved, let us pause here for a minute with it.[408]

The SEC theory does concur with the CS/SAS theory in that Hebbian learning is vital for the development of the representations, depending as they do on the frequency of co-activation of features or items, or their similarity or association value (Grafman 1995).[409] The SEC theory, however, further elaborates the representations and puts them more central in a larger group of cognitive processes, even though these representations rely so much upon prefrontal cortical activations which the original

---

[407] An 'event' in this context is defined by its being thematically consistent although it still can range from a simple motor movement to a more complex event. The SEC is 'structured', as it contains several components in an orderly fashion and with temporal constraints and it is 'complex' because of its consisting of several components that are assembled (Grafman, Sirigu et al. 1993). Here, again, a gradient of complexity is present with less complex SECs being stored in posterior areas in contrast to more complex SECs in anterior areas (Forbes and Grafman 2010). The structure here is understood to be hierarchical, too.

[408] The initial version of the SEC approach did make similar distinctions as the later ones. However, in that version SECs were considered to be only at the bottom of the hierarchy of 'Managerial Knowledge Units'. SECs were then considered to be the 'developmental and phylogenetic precursors' of these MKUs (Grafman 1995). Meanwhile, MKUs do no longer figure in the theory and SECs are considered to become more complex as a result of developmental maturation and especially PFC development.

[409] Indeed, frequent co-activations do generally lead to sparser, simpler representations which in turn can affect subsequent decisions and actions in terms of processing speed and stability (Grafman and Krueger 2008). This concurs with what we've discussed in the previous Part regarding the process or representational redescription as part of learning and developmental processes.

theory considered to be primarily responsible for supervisory processes only.[410] For example, the CS/SAS theory does offer both in its original (Norman and Shallice 1986) and in a later version (Shallice 2002) room for the spontaneous generation of schemas, supported by PFC activations, but limits this capability to the modulation of those lower-level schemas that are employed in contention scheduling. Compared to that, the SEC theory suggests a still larger human capability of establishing event representations at will: "[t]he SEC could be established on the basis of experiencing external events or through the generation of "internal thought"" (Grafman 1995 348). We can expect to see that this human capability of spontaneous schema generation or internal thought and the consequences of such schemas on subsequent behavior is important for our discussion of distal intentions.

Important to note here is that growing expertise – relying on both automatic and controlled processes – affects a range of related processes, including the perception and recognition of novel events. This is demonstrated convincingly in studies of a specific task with regard to event processing, namely event segmentation. Expertise amounts to memory storage of an increasing amount of ever more complex representations, which are modulating subsequent perceptual processes. As a result, increasing expertise corresponds with faster and more accurate recognition of the structure of an action, including its fine and coarse segment boundaries and the hierarchical relations between segments (Kurby and Zacks 2008).[411] Event segmentation research concurs with the SEC theory in that observers primarily recruit prefrontal cortex activation during segmentation tasks, as these activations increase significantly when observers not just passively observe events but have to judge them. However, this research suggests additional recruitment of parietal cortex, specifically for representing temporal features of the events (Zacks, Speer et al. 2007).[412]

---

[410] The contribution of Grafman and others to explanations and theories about action processing has been taken as a further elaboration of the representations involved in CS and SAS processes. Particularly the addition of information about temporal duration and about relative importance to (component) action representation has been recognized as a welcome addition (Cooper 2002).

[411] Event segmentation research is inspired by research of reading comprehension that relies upon readers' construction of situation models in which events and intentional actions play primary roles (Zwaan, Langston et al. 1995). Corresponding to theoretical developments that can be observed in other domains of research of cognition, Zwaan et al. have claimed that the representations found to be involved in reading are in fact employed for a much wider range of mental simulation tasks (Zwaan 2009).

[412] Event Segmentation Theory claims that event representations play an important role in the predictive coding that the brain is constantly doing. Moreover, when errors are detected between predictions and actual observations, the representations are allegedly updated. Apart from its articulation of some additional uses of the representations, the EST has also been presented with an elaborate neural implementation, accounting for many features of event segmentation processes and the features that matter in these. As a result, several cognitive disorders – ranging from Alzheimer's dementia to obsessive compulsive disorder – are partially explained according to this framework (Zacks and Sargent 2010).

Nonetheless, event representations recruit primarily prefrontal areas and these play a central role in the SEC theory, according to which these representations also represent more information than those included in the original CS/SAS theory. This is enabled by the fact that neural implementation of SECs in prefrontal cortex implies their having richer connections to other neural areas. For example, these representations generally contain elements with a modality specific origin.[413] Notwithstanding such an origin, SEC components can still be activated via internal generation or simulation once they have become associated with other and more abstract components. As a result, such an action simulation will still contain properties that are due to its modality specific origin (Grafman and Krueger 2008).[414]

Apart from the modal features that are included in event representations, SECs representations are also considered to have social and emotional features, as these are observed to play a role in action selection processes as well.[415] This again expands their distributed neural implementation. SECs are therefore also associated with activations in cortical or subcortical structures, enabled again by the rich connectivity that neurons from ventral, ventromedial and medial prefrontal cortices have (Grafman 2006 ; Krueger, Barbey et al. 2009).[416] Indeed it is not only the rich connectivity

---

[413] Ridderinkhof et al. present in their review of neurocognitive processes involved in control also an overview of several cognitive processes that are involved in control, like processes aimed at goal-states, anticipation of reward, performance monitoring, error-detection. They also contend that the PFC plays a major role in all these components of control, facilitated by its high connectivity to other neural regions (Ridderinkhof, van den Wildenberg et al. 2004).

[414] In this respect, the SEC theory concurs with accounts of embodied cognition like Barsalou's account of simulators that we discussed in section II.4.1.1 For example, SECs can be considered as representations that underly simulations of a more specific and restricted kind, given that they represent primarily events. Indeed, Grafman and others elsewhere develop 'elators', that is 'event simulators', while referring to Barsalou's work (Krueger, Barbey et al. 2009). However, Barsalou's account explicitly describes how component representations are also stored in sensori-motor areas and not solely in PFC (Barsalou 1999c), which is distinct from the SEC theory.

[415] Reward values have been associated with SECs as well. The explanation of an obsessive compulsive disorder would accordingly be that, due to a neural pathology, an agent may not experience the reward associated with the completion of a hand washing SEC and thus feel the urge to repeat the action corresponding to the SEC (Huey, Zahn et al. 2008). This can be considered an extension of the notion of a component action's 'post-condition' as presented in (Cooper 2002), which was there not associated with reward value experience.

[416] In his account of moral cognition, Moll and others have integrated SEC representations with such emotional and motivational affects, acknowledging the importance of the latter for making moral decisions (Moll, Zahn et al. 2005). Obviously, PFC components do play a role in moral cognition, as is evident from ventromedial PFC activation being involved in social stereotyping, for example (Milne and Grafman 2001). However, other areas also play a role as for example the temporo-parietal junction, which is involved in choosing between selfish and altruistic acts (Morishima, Schunk et al. 2012). Generally, acting according to the results of social (or moral) cognition, though, appears to be dependent upon further motivational contributions. Based upon this, Frijda emphasized the role of emotions for an agent's 'action readiness' (Frijda 1986).

of prefrontal neurons, but also their capability of sustained firing and the fact that prefrontal pyramidal cells are more spinous than those in other areas that supports SEC characteristics (Grafman and Krueger 2008).

With a predominant role for prefrontal cortex according to the SEC theory, it does not present us with a discrete distinction between controlled and automatic processing like the CS/SAS account. It still acknowledges a gradual process of automatization, corresponding allegedly to both the strength of the event representation and its increasingly sparser, economic coding. This process corresponds neurally to a shift from anterior to posterior PFC and facilitates the execution of the representation via neighbouring motor areas (Grafman and Krueger 2008). With such reliance on PFC as areas for storage of SECs, subcortical areas like the basal ganglia are only implied when such an execution takes place (Grafman, Sirigu et al. 1993).[417]

The kludge formation that we associated with expertise is a relatively circumscribed process according to the SEC theory: stronger representations are coded more economically and in more posterior regions (Grafman and Krueger 2008).[418] The CS/SAS theory is in agreement with the relevance of changing activation levels for expertise, with associated action representations becoming ever more complex yet requiring less neural activations.[419] Indeed, both theories agree that as a result, expert action and cognition can become more complex and flexible, yet also be faster than in a novice. In contrast to the SEC theory, the original CS/SAS theory suggested that automatic action does not equally rely upon prefrontal cortical activations.[420] Although the strict separation between CS and SAS processing may have been abandoned, the theory does still make a distinction between the two processes with the process

---

[417] Obsessive-compulsive disorder is explained accordingly with the basal ganglia setting too low a treshold for the activation of prefrontally stored SECs, thus causing their undesirable motor performance (Huey, Zahn et al. 2008).

[418] Apart from this more general prediction based upon the SEC theory, studies of planning and script analysis with patients has suggested that in fact several distinguishable processes like sequence ordering, categorization, and script generation rely upon different prefrontal areas (Sirigu, Zalla et al. 1996 ; Sirigu, Zalla et al. 1995). Indeed, Grafman even suggests a lateralization for integration across events (right PFC) and single event integration (Grafman and Krueger 2008).

[419] Surprisingly, Grafman purports that the CS/SAS model is primarily a processing theory and less interested in the representations involved (Grafman 2006). In light of the extensive computational studies, aimed at simulating empirical evidence of normal and disordered processing of a specific action (preparing coffee) (Cooper 2002 ; Cooper and Shallice 2000), this critique is misdirected. Indeed, these later versions of the CS/SAS model are more explicit with regard to the processing of representations than Grafman's SEC theory is. As the latter is more detailed with regard to possible neural implementations, the integration of both theories promises interesting results.

[420] As mentioned above, an early critique of the assumed separation of automatic and willed control came from patients with frontal lobe lesions which had difficulty in the ordering and the generation of  action scripts – the former task depending upon more anterior areas and the latter on semantic networks that are located in more posterior areas (Sirigu, Zalla et al. 1995).

responsible for automatic action recruiting less frontal areas.[421] To end this chapter on proximal intentions, we will consider neural areas that may underpin the automatic process of action determination.

The process of contention scheduling, underlying automatic action control, was originally suggested to be distinct from controlled action, and to rely completely upon "mechanisms in the corpus striatum of the basal ganglia, often thought to be involved in the selection of actions" (Norman and Shallice 1986 10). Formulated concisely, these mechanisms are responsible for the activation or inhibition of representations that support or conflict with a particular action. In healthy subjects, upon the – willed or automatic – activation of a high-level action representation, a range of related component action representations ought to be activated and inhibited. Later, some specification of these mechanisms was provided.

One mechanism that was mentioned is the striatal dopamine system, which is involved in such activations in the CS processes and of which the deficiency – in Parkinson's disease, for example – is observable in slower initiation of an action. Another mechanism is at work when deficiency of the amphetamine system is associated with failures in inhibiting (component) action representations, observable as the simultaneous performance of multiple, conflicting actions or the repetition of an action that has just been completed (Cooper and Shallice 2000). These disorders are determined by mechanisms that are influencing activation levels in a rather indiscriminate way and thus disturbing the more specific modulation of activation levels of those neural networks that are associated with a particular action representation – whether located in PFC or not.

Cooper later contended that in healthy subjects, it may also well be that the basal ganglia are involved in modulatory processes that lead to selection and inhibition of respective representations even in non-automatic control (Cooper, Schwartz et al. 2005). This contention was partly motivated by evidence that basal ganglia are indeed playing generally an important role in the allocation of neural resources for

---

[421] Research with parietal cortex patients showed that these suffer from problems with specific action script generation tasks, suggesting that relevant representations do indeed not only rely upon frontal areas – even not in controlled processes (Godbout, Cloutier et al. 2004). Other research, particularly aiming at specification of the functions of mirror neuron systems, suggests that representations of relatively simple motor actions for their performance and recognition are stored in premotor cortex (PMC). Patients with ideational apraxia appear to have difficulty in activating the CS scheduling system such that these motor intentions stored in PMC are activated for the recognition of motor actions. It confirms the involvement of CS also in recognition processes and the involvement of regions outside PFC (Rumiati, Zanini et al. 2001). Other research with frontal patients did show that their action representations were not affected like the SEC theory would predict, as it locates these representations primarily in PFC. The effects of the frontal lesions in the patients' performance of verbal and pictorial script tasks were taken to be the consequence of a decreased capability of rejecting wrong alternative responses, which relies upon PFC. This was taken to support the CS/SAS theory and not the SEC theory (Zanini, Rumiati et al. 2002).

the performance of cognitive and motor tasks (Redgrave, Prescott et al. 1999).[422] Even though the strict separation of neural correlates of CS and SAS processes has been rejected and some overlap between the two has been acknowledged, it is argued that in light of evidence – including evidence from patients and from the design of autonomous agents – this account is still presented as a dual-process theory (Cooper 2002).[423]

In closing, it may be useful to return to a model of action selection processes that we referred to in section III.2.2.1. It was discussed there in the context of our discussion of motor intentions and emphasized how multiple motor intentions could arise upon the perception of environmental affordances, requiring the agent to – implicitly, or not – select or decide between those competing action options (Cisek and Kalaska 2010). Indeed, we used the model to confirm our argument that, based upon experience, a form of kludge formation contributes to influencing the action affordance competition such that in experts a sculpted space of actions can be observed. Now when reconsidering this model for sensorimotor control once more, we recognize how it ascribes a crucial role to the interactive activation of multiple, hierarchically structured action representations, as well. The question is, whether the model could be expanded to account also for proximal intentions and in such a way that it integrates some form of secondary or supervisory processing. In that case, it would look similar to the structure of the CS/SAS model.

Indeed, this account of motor intention processing does leave explicitly room for supervisory processes that affect or modulate the affordance competition process which it describes. Indeed, it does so in line with our earlier reasoning in Part II that learning and development tends to employ representations and processes that have already been established earlier, which as a result become even more deeply entrenched. On

---

[422] Based upon interdisciplinary lines of evidence, it is proposed that the basal ganglia have evolved and are particularly well suited for selection problems in general. Moreover, they operate at several levels of specificity of the cognitive or motor options that compete for selection, which is confirmed by the association between their dysfunction and particular disorders in cognition and action (Redgrave, Prescott et al. 1999). Another review equally assigns an important role to basal ganglia activation, in connection with cortical (pre-motor) activations for the generation of relevant action representations (Graybiel 2008). Conversely, for the inhibition of an automatic response an agent must be capable of modulating basal ganglia activations (Aron 2011).

[423] The effect of action familiarity has been studied from other theoretical perspectives, as well. It has been found that familiar action representations are not only benefitting action responses but are also shared with cognitive or semantic tasks. Although not explicitly using a dual-process approach, it is acknowledged that high-level planning does have access to such shared representations as they are capable of activating them, even though the planning activities themselves are both cognitively and neurally distinct (Elk, Viswanathan et al. 2012). Such studies build partly upon the theory of event coding that denies a strict separation between codes that are used in perception and in cognition or action (Hommel, Musseler et al. 2001). Needless to say, that the CS/SAS model is concurring with the notion of shared representations but does emphasize how these may be accessed and employed differently during different task performances.

the basis of their model the authors suggests that other action selection processes are in fact modulating the hierarchically structured selection processes as explained by the model: "[t]he recent evolution of primates is distinguished by advances in the ability to select actions based on increasingly abstract and arbitrary criteria" (Cisek and Kalaska 2010 283).[424] Such controlled selection occurs by way of biasing or strenghtening a particular option, which takes more time than the automatic selection of an action option and does indeed rely primarily upon many forms of PFC activations.

In sum, our discussion of empirical evidence concerning proximal intentions did have a result that concurs in broad terms with the previous philosophical analysis. The latter did suggest that proximal intentions are in fact mediating between motor intentions and distal intentions and in doing so responsible for the anchoring or specification – or blocking, for that matter – of a distal intention in a given situation. Indeed, the distinction of separate proximal intentions has been questioned as such since action determination could allegedly be analyzed in terms of motor and distal intentions alone (Holton 1999). The cognitive neuroscientific model of willed and automatic action which was more closely looked at in these sections concurs largely with that analysis, dividing the task of action control between two distinct types of processing.

If expertise does indeed affect an agent's sculpted space of actions, it must therefore occur through these processes and their interaction. The present chapter has presented insights that proximal intentions are also affected by the modifications that motor intentions undergo as a result of learning and development. To the extent that proximal intentions are to determine action, they can do so mainly by modulating the interactive activations that in a given situation spread throughout a hierarchical network of stabilized – redescribed – motor action representations. The supervisory processes that create these modulations are indeed able to sustain and enhance a coherence that transcends the aims and criteria of the particular situation.

With that, we already touch upon the distal intentions. In the philosophical model presented by Bratman and adopted in modified form by Pacherie, these distal or 'future-directed' intentions are given methodological priority. The implementation of such a priority would amount to a top-down determination of action, undisturbed by bottom-up processes like those which are responsible for motor intentions. In reality,

---

[424] One criterium that is important is of course the expectation of reward. However, there are often several rules applicable to a single situation. Dorsolateral prefrontal cortex may play an important role in favoring one over another rule, with correspondingly different action preferences. This is just one of many potentially relevant criteria, with PFC playing a crucial role in their implementation (Ridderinkhof, van den Wildenberg et al. 2004).

we have seen already that top-down control appears to be limited to modulations of activations of action representations at lower levels. How, then, can distal intentions contribute to long term processes, to an expert's sculpted space of actions? Answering this pertinent question, we will in the next and final chapter of Part III' focus on the issue how an agent can sculpt his space of actions by the controlled (re-)configuration of actions via the explicit articulation of distal intentions. Through such articulation he is both equipping his supervisory processes with rules and criteria that are relevant for proximal intentions and simultaneously modulating the activation levels of the motor intentions that need eventually to be executed.

# 4 DISTAL INTENTIONS: GOVERNING THE INTENTIONAL CASCADE?*

Based upon ample professional and personal experience, our expert singer knows how to use his voice and body optimally when he tries to seduce Zerlina. A sculpted space of actions allows him to flexibly and without much thinking use the stage-props at hand while interacting with her – his cape and sword, perhaps a scarf that Zerlina is wearing, and so on. Proximal intentions help him to anchor his Don Giovanni performance in the context of this particular stage and specific interpretation, while implicit motor intentions facilitate at a finer grain his making appropriate manipulations of the props and convincing vocal expression of the score. Given the constraints that determine his sculpted space during a performance, he need not be constantly reflecting on his action options. He need not to put effort in looking away from the female singer and up to the heaven since he is not playing the celibatarian Saint François nor must he suppress aggressive or even suicidal tendencies when handling his sword since he is not impersonating Peter Grimes. Instead, he presents a juvenile and heroic character with the corresponding motor and vocal performances without requiring much awareness. That is, if he is to perform as he is used to do, as he has seen others doing and as he beliefs it should be done.

On stage as in everyday life, however, things do not always go as they normally do. A stage director comes along with a revolutionary interpretation, a conductor wants to show the score in a surprisingly new light, or our singer himself is not convinced that he should just repeat his earlier performances. As a result of either of these interventions, Don Giovanni is now required to behave shy, not take his sword but a pencil in the hand and sing with a somewhat throaty voice. This would require our singer to exert careful control of his performance, demanding perhaps some extra rehearsal time as it implies that he needs to suppress habituated actions, to anchor his role anew in the specific situation and then to practice these new and alternative actions so well that he can perform them and meanwhile attend to the conductor, intonate properly with Zerlina, and so on. Being a professional singer, he has established a distal intention to comply with a new direction and the actions that are determined by it, now working hard to specifically adjust his proximal and motor intentions where necessary. A reconfiguration of his sculpted space of actions is required for that.

But sometimes an experienced agent is forced to more than modify his space of actions in a habituated situation. It is not hard to imagine a stage director who wants to

---

* On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively. Fig. III is particularly relevant as a representation of the main contents of section III.4.

portray Don Giovanni as an immoral, egocentric rapist and asks our singer to behave accordingly, to attempt to strangle and violently undress Zerlina, to walk around half-naked and to sing with a harsh and chilly voice. If so, we can imagine our expert singer not willing to conform to this interpretation, as he may have fundamental objections against it. He may have artistic objections against this excessive interpretation, may be too prude or insecure to undress himself, may want to avoid harming his voice or may fear to alienate his wife from himself when she would see his performance, for example. In these cases one or more distal intentions that happen to govern over a longer period of time his personal and artistic life can make him decide not to accept the role this time.

Perhaps, he at first thought that things would turn out not so bad, only to find out during the first rehearsals that the direction was even more ridiculous than he thought, that he indeed felt strongly embarassed or that he disliked singing as he was required to do. Upon having experienced how he would need to specify and anchor his actions during a scene rehearsal, then, he would be certain of his decision not to comply with this interpretation even though he usually let himself be guided by a director. He would than know that it would demand him to not only block the application of some of his distal intentions in this exceptional case but that it would force him to transgress the limits of the space of actions associated with these intentions. More than just a matter of having to control and adapt his habituated actions he would need to perform such that he would feel like betraying himself. At least, that is how the situation appeared to him after sufficient reflection upon his first response and the experiences during the first rehearsal. Indeed, even an extensive discussion with the director and a colleague friend could not deflect him of his decision, of which he now is certain.

This episode from an artist's life shows us how an agent's performance is constantly determined by the simultaneous influence of intentions that operate at very different levels of grain. It also demonstrates how difficult it can be to eventually settle for a particular action when it seems impossible to act in such a way that coherence between the three levels of intentions obtains. It may take an agent a while to at first set aside his feeling of uneasiness, then to experience personally his embarassment which strengthens his initial feeling and finally to decide to stand behind his objections. Unfortunately, this process is often necessary as there are no strict laws or rules that determine which motor, proximal and distal intentions specifically cohere together and how, and which do not.

Indeed, we've observed that the the intentional cascade model does leave the three levels of intention a certain leeway to each other, while it simultaneously provides top-down and bottom-up interactions. For example, we've discussed in section III.2.1

how an experienced agent will have established a set of repeatedly activated motor intentions that then respond differentially to specific environmental affordances with the prepotentiation of corresponding action responses. Given such prepotentiations, it is likely that in a given situation the anchoring of a distal intention will take place by executing one of these prepotentiated motor intentions instead of developing a novel one. That is, if the agent has not formed a distal intention that specifies an uncommon course of action that requires him to withhold this habitual response and to do something else instead, for which the proximal intentions discussed in section III.3.1 are involved.

Distal intentions play in both cases an important role, as they govern in some sense the development of an overarching action plan and are also involved in withholding an implicated habitual and choosing an alternative action instead. The latter could be necessary if the habitual action turns out not to fit an action plan, or if an agent is even modifying his action plans. Accordingly, distal intentions have been said to have three main roles: "as terminators of practical reasoning about ends, prompters of practical reasoning about means and plans, and intra- and interpersonal coordinators" (Pacherie 2008 182). The first role seems natural, as ends – especially long-term ends of a planning agent – transcend the specific environmental and corporeal conditions which we found in earlier sections to figure prominently in proximal and motor intentions. Therefore, ends that lie more in the future require other capabilities than involved in those latter intentions. The second role of distal intentions does refer to the fact that, still, for its realization an agent's practical reasoning needs subsequently specify necessary means and plans. The third and final role is perhaps the most difficult one, as coordinating one's actions can be a quite complex task.

Indicating this complexity, Pacherie distinguishes between three different forms of consistency which rely on distal intentions that are to be considered. First, the component actions involved in a distal intention should be internally consistent and not cross each other. Second, an agent's distal intention should not contradict what he believes and thinks about the world, should be externally consistent. Third, global consistency is required for a distal intention to be integrated with his other action plans (Pacherie 2008).[425] The question presents itself, how consistency across such different domains or contexts can be established, covering actions, beliefs and plans? What representational format lends itself for fulfilling this role?

---

[425] Bratman does not make precisely this threefold distinction. He does discuss consistency constraints, but seems to limit these to what Pacherie calls the internal and external consistency constraints. However, he does then also acknowledge the "demands of consistency on my total web of intentions", which amounts to Pacherie's global consistency (Bratman 1987 32).

This question is not completely new as we've seen in the context of motor and proximal intentions that they require integration of specifics regarding muscle movements, situational affordances and about action goals. Different forms of information can be represented and processed by the brain such that they become associated with each other even though their differences – and the different processes in which they are involved – remain intact. In the present context, however, Pacherie claims that a shared representational format is necessary if distal intentions are to play their coordinating role. For she contends that for distal intentions to be able to function as consistency enhancing coordinators, they rely upon a "network of inferential relations among intentions, beliefs and desires." According to the argument, it is the conceptual nature of distal intentions that allows such a network between intentional states to be developed: "[t]heir sharing a common conceptual representational format is what makes possible a form of global consistency, at the personal level, of our desires, beliefs, intentions and other propositional attitudes" (Pacherie 2008 184). Are we indeed to assume that distal intentions are fraught in a conceptual representational format and are to exert their determining role in the intentional cascade based upon a format that is potentially very different from the format of motor representations?

Within the context of the intentional cascade, this issue is not explicitly addressed. Elsewhere, however, Pacherie admits that the conceptual representation involved in a distal intention must be connectable or even convertible into appropriate motor representations if it is to be executed by way of an action. Even though, as we noted earlier, motor intentions are cognitively impenetrable they can become associated with the conceptual contents of intentions depending on an agent's experience with both the motor movements and the relevant concepts (Pacherie 2011).[426] Here again, she leans on the work of Jeannerod, who has demonstrated that there is an equivalence between motor preparation and motor imagery and who "suggests that the same general framework used for simple object-oriented actions remains applicable to higher-order representations encoding long-term action plans" (Pacherie 2000 413).

Important aspects of that general framework are that actions are represented in a distributed way in the brain and that the representations are organized partly in a hierarchical way. Moreover, these representations are used more or less equally for

---

[426] In a chapter written together with Haggard, Pacherie clarifies that forming a prospective intention – the term used there for distal intention – does in fact amount to the activation of mental representations of relevant environmental cues, for example. Given the association of such representations with relevant motor representations, a distal intention can indeed determine activation of non-conceptual representations as well (Pacherie and Haggard 2010). Supporting this analysis, the authors refer to the effectivity of implementation intentions in governing an agent's future actions (Gollwitzer and Sheeran 2006).

the execution, imagination and observation of actions, in which case representations are assembled from various motor action components. Jeannerod refers to that process as 'motor simulation' (Jeannerod 2006). However, he acknowledges that his account of motor simulation and its prevalence in various cognitive processes focuses on representations of motor actions only, leaving out many relevant aspects of the representation of actions – like the objects or tools that are involved in many actions. Furthermore, motor simulation in this account has a limited validity with regard to complex, temporally extended and hierarchically organized actions and to actions which require prior knowledge about the context and agent that are involved (Jeannerod 2006).[427] If the approach is to be useful in this context, we therefore may need to take on an expanded version of it.

The reader may recall that in chapter II.4 ff. we discussed another simulation theory, that had a much larger scope as it was based upon the assumption that: "simulation constitutes a central form of computation throughout diverse forms of cognition, where simulation is the re-enactment of perceptual, motor and introspective states acquired during experience with the world, body and mind" (Barsalou 2009 1281; cf. Barsalou 2008). According to this theory, a simulation offers an assembly or reconfiguration of component representations that stem from previous sensory, motor or cognitive experiences and that have been stored in a highly distributed fashion across the brain. Unlike Jeannerod's motor simulation theory, this simulation theory does not exclude domains of experience from being integrated in such a simulation.

On the contrary, although we found Barsalou's simulation theory emphasizing the modal origin of 'perceptual symbols', it does acknowledge that humans even have the capability of producing abstract concepts or mathematical concepts by focusing attention on or reassembling particular components of stored – modal – representations for which language is important (Barsalou 1999c). In much the same vein humans can employ language for configuring the representation of a potential situated action by engaging not only the areas in which relevant sensory and situational information is stored, but also areas that are involved in motor control (Pezzulo, Barsalou et al. 2011).[428] What we've learnt in Part II about the phenomenon of the re-description

---

[427] It appears that Pacherie has not really faced these limitations of Jeannerod's account when developing her own. In another article on 'The content of intentions' she offers a short reflection on the perception and processing of 'scenario-content' and 'protopropositional content' and connects this to a reflection on the representation of basic action concepts (Pacherie 2000). What she does not offer is an analysis of how especially distal intentions can contain representations of a variety of contents, objects, and actions.

[428] This simulation theory does not only offer a theory of grounded concepts but also leaves room from a reverse influence, in which case a simulation results from focusing attention on a particular concept, for example (Barsalou 2009; Barsalou, Cohen et al. 2005).

of representation is therefore something that is not restricted to early stages in child development or learning, but is facilitated by language in a more general sense.

Important to note with regard to our interest in an agent's sculpted space of actions is that one may think of a simulator as not producing just a single simulation of a concept or situation. Instead, it is argued that "the space of possible simulations within a simulator is as a space of Bayesian priors, with some simulations being more likely than others" (Barsalou 2011 191). This conception of a sculpted space in which some options are more probable than others concurs with the theory regarding interactive activations of hierarchically organized action representations (Cooper and Shallice 1997 ; Norman and Shallice 1986) that was found to be helpful in explaining the functional roles of proximal intentions.[429] However, the simulation theory does even extend the latter theory, as it is explicit about the fact that all kinds of components – modal and amodal or abstract - of a simulated action are stored in a distributed way throughout the brain and can be configured through imagination or speech and consequently influence future cognition and behavior (Barsalou 1999a). A consequence of that extension is that it also contends that a simulator can be supported by a wide variety of neural areas depending upon the introspection involved (Barsalou 2003) or whether novel and unrealistic situations are being simulated (Barsalou 1999c). With all these areas and processes involved under the influence of previous experience, there is a distinct advantage for experts over novices. In experts a sculpted space, containing more and richer simulations, will be activated in specific situations, subsequently facilitating in a flexible way a fast and adequate action response (Barsalou 2009).[430]

With these preliminaries in place, we may ask in what sense this process of developing a simulator, for example by formulating a distal intention, is subject to constraints or not. For in line with our arguments so far, it is to be expected that a distal intention can play its role in coordinating actions both at a personal and an inter-personal level best if such constraints have been developed in the context of distal intentions as well. If this were not the case and the constraints corresponding to

---

[429]  Again, as was the case in the comparison with Jeannerod's simulation theory, Barsalou's theory of simulation aims to account for the infinity of complex and recursive configurations that can be made available through selective attention and language, among others (Barsalou, Simmons et al. 2003).

[430]  As has been noted earlier, it is debatable to which extent experts may have a disadvantage in coming up with novel and creative solutions given the presence of a sculpted space which may predetermine their responses. If, however, creativity is taken to depend upon the recognition of relevant dimensions of a problem space in order to develop a useful novel solution  – as Boden has argued in her influential (Boden 2004) – then we can expect those to come from an expert rather than from a novice. This is not to deny that an expert may have more difficulty in suppressing solutions that have become habituated, whereas novices are not burdened with that difficulty. This is captured nicely in the saying: "Children love kitsch but make art whereas adults love art but make kitsch" – as adults often prefer modern art but have hesitations to express themselves as directly and unpolished as children do.

a sculpted space would only emerge from motor and proximal intentions, the question is whether such constraints would be sufficiently coordinating actions and would be sufficient in providing maximal coherence and concistency in an expert's actions.

Let us start by scrutinizing more in detail distal intentions's role in determining and coordinating actions is articulated by Bratman and then turn again to empirical findings that bear upon this role.

## 4.1  A philosophical analysis of distal intentions

In section III.2.0 we referred to Grice's and Bratman's exercises in creature construction in which both drew increasingly complex creatures as a way of demonstrating what adding further capabilities implies for their functioning. The creatures that Bratman presents develop already in their fourth – of eight - stage planning capacities, under the pressure for coordination and organization both of its individual actions and its interactions. Irrespective of moral issues it is the possibility that any action can be counter-productive or costly in multiple senses, that planning agency was found to be required to contribute to increasing coherence. However, planning turns out to be more complex than at first realized in the planning Creature 4, as successful planning requires ever more conditions to be fulfilled. In successive steps, therefore, planning agency is expanded such that it involves the capability to develop a hierarchical structure for its intentions, a structure that is partly determined by a reflective valueing. Through this reflective valueing it is governing not just its actions but also its intentions and desires, yielding self-governing policies with which it can not only organize but even justify favoring some over other of its desires and intentions (Bratman 2006b, ch. 3).

This suggests that an agent equipped with a self-governing policy must constantly engage in self-reflective deliberation about his intentions for action. However, this would contradict our observation earlier, in the context of motor and proximal intentions, that the development of a sculpted space of actions allows an agent to act in most situations fast and flexibly without demanding costly deliberation and reconsideration. Consequently, the question is whether and if so, what, a self-governing policy can contribute to this space and under what conditions an agent should spend time and resources on deliberatively further sculpting his space of actions. On the basis of the reflections on distal intentions above we can already predict that they probably must rely upon a rather complex set of capabilities, with equally complex neural underpinnings. The latter, though, will be our topic only later.

### 4.1.1  Closing the gap between distal and proximal intentions and sculpting the space of actions

In most situations, multiple affordances for a motor action are available for an agent and equally more than just a single proximal intention can be realized. Constraints could be helpful in assisting an agent to escape from being paralyzed and unable to choose which action to perform or, conversely, from performing actions that are counter-productive and incoherent with his other actions. These constraints should sculpt his space of actions such that in each situation the probability for most actions of being executed is decreased, while eventually a particular action option – constituted by a certain configuration of action components – is executed. As we have seen earlier, experience contributes a lot to the development of such constraints via processes like information chunking, association formation between perceptual information and motor responses, and kludge formation that facilitates habitual action in common situations. Over time, these constraints facilitate processes like contention scheduling and the supervisory control of action to proceed relatively fast and flexibly, with minimal taxing of the agent's cognitive resources.

However, if his sculpted space of actions were to be only constrained by way of his experience and the corresponding motor and proximal intentions, the agent would face two difficulties with regard to the distal intentions that he has developed or will formulate in the future. First, if his distal intentions would not be somehow integrated in his sculpted space, their practical anchoring, specification and realization through associated proximal and motor intentions would likely be impeded. If each distal intention would be new to an agent's sculpted space, such anchoring, specification and realization would always require the top-down establishment of novel associations with proximal and motor intentions. Given bottom-up influences, like the prepotentiation of motor intentions as a response to perceived affordances, the execution of a completely new distal intention might then also require the inhibition of motor and proximal intentions that figure prominently in an expert's sculpted space of actions. Second and related to this, developing distal intentions does not yet protect an agent from committing incoherencies. Particularly because of the agent's capability to execute complex intentions that imply component actions across time and space, potentially engaging other agents as well, his sculpted space of actions may not yet be sufficient for helping him to avoid incoherencies and creating coherency. As Bratman's account of planning agency is particularly aiming to elucidate how organization and coordination of actions is established, we will consider how he thinks we can aim to close this gap with distal intentions.

Now distal intentions – Bratman's future-directed intentions – are meant to answer

two human needs, the first of which points more to a limitation, the second rather to a capability. For it is because of our limitations in terms of time and cognitive resources that we cannot always rely upon deliberation during a particular situation to see whether it is suitable for developing a proximal intention, and if so what. Apart from this limitation, it is our capability of realizing complex goals that requires additional intrapersonal and inter-personal coordination. Distal intentions are involved not only in planning such complex goals but also in the associated coordination role (Bratman 1987). It may be tempting to think that an agent should then be as specific and concrete as possible when forming an intention at a place and time that is 'distal' to situations in which concrete actions have to be performed. However, such comprehensive planning is practically impossible and would deny the unpredictabilities of reality, like the unpredictability of another agent's behavior. Instead, planning agents' plans are usually incomplete: "we typically settle on plans that are *partial* and then fill them in as need be and as time goes by" (Bratman 1987 3, italics in original). This leaves distal intentions in an ambiguous position, being involved in planning and coordinating, while leaving much of those functions to later moments, when proximal intentions have a crucial role in such filling in or specification.

The consequence of that peculiar position is that a distal intention does enforce an agent to further deliberate upon his intention, now or at a later stage. Fortunately, it does so not without offering resources to constrain that process. Depending upon the comprehensiveness of a distal intention, he must start to specify and plan the means to reach the intended end at an early stage: intending as a child to become an expert singer requires longer preparations with many more intermediate actions than intending to be on time in the opera house tomorrow. In both cases, however, the distal intention does provide some resources for constraining the necessary deliberations, according to Bratman: a distal intention can also play a role as a 'filter of admissibility for options' and offers 'standards of relevance for options', which makes the admittance of options and a subsequent selection between them more feasible (Bratman 1987 33).

This role is especially played by those distal intentions that are integral to a characteristic component of this account of planning agency and probably also quite relevant for our sculpted space of action: a person's policies. An agent typically has such policies, which consist of a habitual way of acting in or responding to recurring situations. Our expert singer, for example, may have as a policy to avoid smokey places the night before a performance. Such a policy then constrains and guides his further deliberation about distal intentions, like his joining his colleagues after tomorrow's show at a party.[431] Having a general policiy is therefore useful for three reasons, according to Bratman's account. First, it solves the problem of our limited time and

resources as it has his deliberation influence his intention formation at a later moment. Second, it simultaneously helps in interpersonal and intrapersonal coordination. Finally, it invites an agent to focus on and decide about comparable situations as it "may sometimes be easier to appreciate expectable consequences (both good and bad) of general ways of acting in recurrent circumstances than to appreciate the expectable consequences of a single case" (Bratman 1987 88). The fact that similar situations occur repeatedly allows a planning agent to alleviate his future deliberative efforts, it seems, and nonetheless increase the consistency in his actions.

What such a policy brings about, then, is constraining an agent's distal and – with that – proximal intentions by constraining the deliberation that is involved in the formation of these intentions. Consequently, in a case when a distal and proximal intention are based upon a policy, there is not 'full-blown deliberation' of all possible action aspects involved (Bratman 1987 90). Such deliberation is no longer necessary as in recurrent circumstances the relevant aspects of a situation and the corresponding action consequences can be predicted based upon previous experiences. The policy is in such a case *circumstance-triggered*, Bratman argues, and we may add that the corresponding, policy-based intentions do not require further deliberation (Bratman 1987 88, italics in original). With regard to such proximal intentions, we learnt above that in an exceptional case an agent may feel forced to for once block the application of his general policy – for example when he does not buckle up in an emergency or when our singer decides once to deviate from his policy and joins his colleagues tomorrow at a birthday party at a smokey party. Recognition of the exceptionality of the current situation, based upon great expertise with recurrent normal situations is required to ignite the deliberation that is otherwise likely superfluous.[432]

This observation that a general policy may in exceptional circumstances be blocked – not rejected nor reconsidered, but just blocked – points to a fact that has not received yet much attention: the challenge for a planning agent to coordinate and organize multiple intentions and policies. Sure enough, we've noticed that a third form of consistency was included in the intentional cascade model, since global consistency

---

[431] With regard to the proximal intentions, we already discussed the possibility of blocking the application of our general policy – for example to buckle up during driving. In that case, it is not the policy that we reject or reconsider, but only its applicability in an emergency situation (Bratman 1987). The singer may similarly decide once to deviate from his policy and distally intend to join his colleagues tomorrow at a birthday party at a smokey party.

[432] Indeed, Bratman argues that based upon our habits and propensities we "take notice of certain sorts of problems but not of others – to treat certain aspects of the environment as salient" (Bratman 1987 65). Adequate expertise, so we argue, does also imply that an expert notices when the environment presents a situation that potentially outwits his habits and skills. More below, we will argue that imagination and narrative may be a means of preparation for such situations to the expert.

referred to the integration of a plurality of action plans and distal intentions (Pacherie 2008) or the consistency of his 'total web of intentions' (Bratman 1987 32). One would expect that this consistency does not require constant deliberation, too, but is maintained by the kind of dispositions that we've been considering in this section.

### 4.1.2  Hierarchy and stability of a planning agent's web of intentions

Not buckling up in an emergency situation or entering a smokey environment for a specific celebration implies that the agent's policy is blocked because another policy or distal intention has now been given priority: saving someone's life or affirming a friendship. In such cases it becomes apparent that an agent usually has adopted multiple policies, which can conflict with each other. This requires an agent – striving to avoid costly and counterproductive actions - to arrange his policies such that chances for conflict are diminished. The commitment to his policies cannot be equal but must therefore be differential.[433]

A planning agent is required to coordinate his policies with each other, which does require him to reflect upon them and systematically arrange them. In line with our arguments so far, it is to be expected that the arrangement of his policies will have a hierarchical structure – in addition to the hierarchical structure of his actions and plans. Only with such an arrangement will he be able to respond to coherently yet flexibly to both common and exceptional situations, according to his own intentions and plans. This is no longer just a matter of rational reflection or deliberation, but also a matter of 'valueing'. Assigning values to his policies and plans in such a way that priorities result can help to avoid or solve conflicts or inconsistency: "an autonomous agent not only governs her actions but also governs the practical reasoning from which those actions issue" (Bratman 2006b 164).[434]

What is required in addition to an agent's planning capacity is a 'motivational hierarchy' that can guide his planning or policy-making whenever necessary. Interestingly, Bratman argues that such a hierarchy cannot rest exclusively upon rational argument or deliberation, since it involves the assignment of values to the policies or higher-order intentions that are used in reasoning and deliberation. The

---

[433] Frankfurt argues even that a person can only be genuinely free and willful, if he has identified with certain higher-order 'volitional necessities' or ideals. In that case, he must no longer reflect and reason about them, even though these guide his choices and autonomous actions: "Volitional necessity constrains the person himself, by limiting the choices he can make" (Frankfurt 1999 113).

[434] Such conflicts are different from those that have to do with deliberation about means towards ends or reasoning about ends. More general conflicts stemming from an agent's commitments are meant here: "human agents are complex, and in many cases of interest, there is conflict among relevant practical attitudes" (Bratman 2006b 260).

hierarchy puts in place not just policies but higher-order policies or 'self-governing policies' that "say which desires are to have for the agent what we can call "subjective normative authority"" (Bratman 2006b 210). Surprising as it may be, we find here an argument for there having desires or 'conative attitudes' this normative authority and not only rational justifications.

This motivational hierarchy, that needs to be in place for a self-governing planning agent, is subject to a similar limitation that is relevant for his intentions and policies.[435] Being still a temporally extended agent with highly limited resources and capabilities, he must avoid costly and counter-productive actions or reflections, which now implies that he should also avoid as much as possible reconsideration and revision of this hierarchy. If not, the 'system of self-management' would break down as bodily appetites, emotions and other motivating attitudes would "challenge and/or diverge from, our commitments to weights and other forms of significance" (Bratman 2006b 241). So even if his 'satisfaction' - which Frankfurt expects an agent to have with the structure of his will in order to be autonomous (Frankfurt 1999) - would be tempted, the agent should refrain from reconsidering or deviating from his motivational hierarchy.[436]

Interestingly, here again we find that it is reasonable for an agent not to constantly reconsider his intentions, reason about his policies, or reevaluate his underlying motivational hierarchy, since he would then lose time and resources and would probably perform counterproductive actions. If, instead, he would have established a sculpted space of actions that would also be constrained by his distal intentions and policies, this would "support some sort of defeasible, default presumption in favor of following through with one's prior intentions and policies" as is being called for (Bratman 2006b 276).[437] This raises the question how such higher-order constraints can be included in this sculpted space. Now that we've gathered support for such a sculpted

---

[435] Concurring with the increasingly complex nature of the creatures in Bartman's exercise in 'creature construction', he here argues that "a basic pressure for conative hierarchy derives from what is for human agents a pervasive practical problem of self-management" (Bratman 2006b 219).

[436] In his reply to Bratman, Frankfurt reconsiders his formulations concerning the role of higher-order intentions and the identification with these intentions or the endorsement of such intentions by an agent. Frankfurt regrets these formulations since they suggest that the agent must attach a value to them or treat them as providing justification for his intentions, which is not what he meant. Indeed, in contrast to Bratman, Frankfurt doubts whether deliberation and practical reasoning are so important for human agency, partly since animals can have a form of agency without them (Frankfurt 2002c). We concur with Frankfurt's reservations insofar that we argue that deliberation and practical reasoning need to be complemented with other capabilities like imagination or simulation, that may to in some form be shared with animals, to. Comparative studies suggest that some animals are indeed capable of simulation or 'mental time travel', as is being discussed in (Suddendorf and Corballis 2007) and its commentaries.

[437] Moreover, it is not just the presence of intentions and policies that support non-reconsideration. There are also many nonreflective cases of non-reconsideration, due to 'certain underlying habits, skills, and dispositions' (Bratman 1987 60). Naturally, such habits, skills and dispositions are often the result of previous deliberation and reflection, as we've argued earlier and our example of the expert singer has testified. In addition, we will now also consider a role for simulation and narrative.

space, let us see whether this account of planning agency also offers suggestions for expanding it at this point.

### 4.1.3  A role for the imagination in weaving one's comprehensive web of intentions?

During our discussions of how motor, proximal and distal intentions and self-governing policies might be involved in an agent's sculpted space of actions, we have also noticed that their contributions to this space are to a large extent an effect of the habits, dispositions and attitudes that they lead to. That is, rational deliberation and argumentation play a limited and not an exclusive role in planning agency. There are in particular two functions, for which other resources might be useful, for developing a more comprehensive web of intentions and that perhaps in turn contribute to his sculpted space. First, it may be useful for an agent to gain some form of preliminary or preparatory expertise in anchoring and specifying his distal intention in various situations. A way to do this could be to imagine oneself in executing multiple proximal intentions that all appear to implement that distal intention. Second, rational arguments may not (alone) suffice to an agent's development of self-governing policies, without which he would have difficulty in giving priority of some of his policies over others. The necessary motivational hierarchy and values may therefore need another source or support. Here again, imagining oneself in a situation in which distal actions based upon some of his policies – with which an agent is otherwise satisfied - might lead to inconsistencies,  could be a useful endeavour and challenge him to assign priorities. Unfortunately, however, the account of planning agency and the corresponding intentional cascade seems not to offer such resources, complementary to those that we have considered.

Now Bratman does acknowledge that an intention cannot always be decided upon in a straightforward way, even not against the background of an agent's established web of intentions and beliefs. For notwithstanding this background, there may still be acceptable and relevant alternatives available for a particular intention, each having different reasons for and against it. A first step is that an agent has at his disposal a further 'screen of admissibility' that helps him to rule out some intentions when they are considered against the background of other intentions. For an agent should "avoid functionally incompatible intentions" (Bratman 1987 162), that is: he should avoid having dispositions and intentions such that their simultaneous realization is impossible.[438] The question is, however, how he can realize such global consistency when at the same time his distal intentions can be partial and be left open for future specification.

To be sure, planning agency does involve some anticipation of future consequences of intentions. For example, in his exercise in creature construction it is already with Creature 4 that a structure of 'anticipated future regret' is implemented which should help it to stick with its execution of a plan or policy (Bratman 2006b): although future-regret is here merely assisting the creature in resisting temptations to deviate from its plans, it ascribes to it capabilities to envision consequences of planned and deviating actions and its own responses to it.[439] Moreover, the anticipation is elsewhere clarified as not merely referring to a counterproductive action but rather being the "anticipation of a breakdown in the cross-temporal coherence of this temporally extended agency" (Bratman 2006b 277). In both cases, however, the consequence drawn because of this anticipated future regret is quite modest: it is taken to support the importance of an agent's non-reconsideration of his stable plans and policies. It is not taken to force upon the agent active investigation or imagination of whether future situations may give rise to inconsistent actions or to deliberate whether his partial plans and policies might lead to inconsistencies when further specified and filled in.[440] So although

---

[438] In a later discussion with Frankfurt's notion of an agent's identification with his desires, Bratman develops the idea that an agent must be 'satisfied' with his intentions and policies, which implies that such a policy "needs to be free from significant challenges from other relevant higher-order policies" (Bratman 2006b 83). In that context he positions his account between a Humean theory of action as being caused by desires and will on the one hand and a Kantian theory determined by universal principles. Both theories, however, do not offer the resources of the problem we're discussing here: how an empirically plausible theory of action can account for an agent's actions being consistent at several levels of specificity.

[439] The role of future regret in Bratman's account of planning agency resonates interestingly with Aristotle's peculiar argument in his Nicomachean Ethics (book II.3) that we can partly judge the correspondence of an agent's character with his action – and consequently its voluntariness - by observing whether he responds with regret or other emotions after the action. Although for some interpreters this argument has been a reason to consider the other, Eudemian ethics as later or superior – (Kenny 1979) – we consider it as an empirically plausible addition to the reflection on voluntary action. Ricoeur, simiilarly, ascribes importance to an agent's character which include 'aspects of evaluative preference' and connects this notion also with the identity of an agent (Ricoeur 1992 122).

[440] Somewhat related to this lack is the absence in Bratman's work reference to the so-called 'frame problem' which arose in the context of AI. The frame problem raises the question how much background knowledge must be implicitly or explicitly represented for an agent – artificial or not – to be able to perform a particular task, including knowledge about the stability of properties of objects in the world, etcetera. Dennett refers to Minsky's work on 'frames' and Schank's work on 'scripts' as a way to tackle the problem (Dennett 1987). We have emphasized the importance of such representations for our agent and of their contribution to a stable sculpted space of actions. In our discussion of narrative, we will also note how tradition can be the origin of complex action configurations, which helps an agent to circumvent somewhat the frame problem by assuming their efficacy. Bratman, on the other hand, focuses in most of his work on the framework that an individual agent over time erects as a background for his own action planning. 'Shared valuing and frameworks' are treated mainly in the context of a particular joint action or policy, not in order to solve this frame problem (Bratman 2006b). In an 1992 article, Bratman argued that intentions must be context-independent, even if the decision upon which they are originally based are context-dependent. Only when context-independent, intentions can fulfill their organizing and coordinating role over multiple contexts (Bratman 1999b). Contextual knowledge as implied by the frame problem is not at stake here, either, even though it is assumed silently to support this role for intentions to play.

anticipated future regret does signal the agent's imaginative capabilities and even attaches a functional role to it, the exercise of such capabilities of is not included in the account of planning agency as presented to us.[441]

There are a few remarks of Bratman's that hint in that direction, although they are concerned merely with weighing alternative actions against each other. Nonetheless, they do refer to an agent's capability and sometime necessity to use his imagination to solve a difficulty in determining his action, as there are no obvious reasons for or against it. For it is acknowledged that with regard to weighing alternatives against each other, sometimes an agent must use his imagination in order to settle the issue and to draw his preference: "[t]he agent will attempt to weigh conflicting reasons by rehearsing in imagination just what would be involved in acting on one or the other of those reasons" (Bratman 1987 59). However, although it seems that reasons are weighed here not in terms of their rationality nor in terms of their calculated utility, it is left undetermined what criteria are used. Instead, weighing occurs in terms of the imagined actions that are associated with those reasons, which may require "procedures of "dramatic rehearsal (in imagination) of various competing possible lines of action"" – as Bratman says while quoting Dewey (Bratman 2006b 150).[442]

Unfortunately, Bratman does not continue this line of thought nor does he elsewhere articulate what other psychological functions or capabilities are needed to complement his account of planning agency.[443] Neither does he consider what Dewey had to say about this dramatic rehearsal. The latter did further articulate dramatic

---

[441] This absence may have to do with Bratman's conviction that although decisions for action are made against a certain context and therefore context-relative, the resulting intentions are context-independent as they must be "compatible both with the agent's relevant beliefs and with the agent's other intentions" (Bratman 1999a 32). It is precisely their role in the comprehensive web of intentions and beliefs that intentions should be independent in that way. This is different from the narrative approach by Ricoeur, as we will see, which even entails that articulating a narrative will time and again affect even the agent's identity – or selfhood (Ricoeur 1992). As a result, the background against which intentions and promises are made are to some extent always in flux, for example.

[442] Interestingly, although imagination is nowhere explicitly addressed in Bratman's works as far as we have been able to find, it is referred to twice – both with regard to this same quote from Dewey. It may not be incidental that the quotes differ, though. The 1987 quote runs: "assess the rationality of [agent] S in employing such procedures of "dramatic rehearsal (in imagination)" in his deliberation" (Bratman 1987 59), which suggests that the imagination is here employed in support of rational assessment. A different emphasis is put on the quote in the 2003 article, that has been quote above. In this more complete version, imagination is explicitly devoted to a 'rehearsal' or – as we would say – simulation of possible actions. Based upon such simulations, an agent may come to commit himself to a line of action for which it cannot be definitively reasoned, but which he does also not just want to arrive at through 'unreflective, brute picking' (Bratman 2006b 150).

[443] Oddly enough, Bratman comes close to attending to imaginative rehearsal when he considers the role of 'anticipated future regret' that an agent might experience when he imagines himself to give in to a temptation which makes him deviate from his normal policy and thus breach his agential consistency (Bratman 2006b 286-287).

rehearsal of lines of action, as "an experiment in making various combinations of selected elements of habits and impulses, to see what the resultant action would be like if it were entered upon" (Dewey 1922 190). In this rehearsal or simulation – as we will refer to it - , external objects and agents are said to also play their part, all elements together providing the agent some sense of meaning associated with lines of action. Lines of action, we may paraphrase, that essentially contain imagined recombinations of components belonging to his repertory of habitual actions and intentions.

Let us, unlike Bratman, carry this line of Dewey further and consider whether some form of rehearsal – dramatic or not - can contribute to an agent's sculpted space of action in a way that enhances his coherent self-governing agency.[444] For example, can an agent experience with such rehearsed recombinations or reconfigurations in a limited experience what realizing a distal intention might amount to, or how certain intentions or policies might clash with each other in a given – perhaps exceptional – situation? More important even, would that rehearsal-dependent experience not only help him to prepare for future actions but also have an immediate impact on his sculpted space of actions? If such imaginative rehearsal or narrative simulation indeed has such an effect then engaging in it may be meaningful for his agency even if it is not immediately recognizable as such. Let us now, in addition to the rational argumentation and deliberation about actions discussed in this section, consider in the next section whether narrative simulation can play a relevant role and if so, how.

### 4.1.4  Narrative simulation as an additional resource of establishing one's agency

Especially when distal intentions are concerned, the account of a self-governing planning agent assumes that he is capable of considering an action which might unfold in a future situation. Meanwhile, we have learnt that this is no small feat: for although his sculpted space of actions may help to constrain the space of action options that he will consider and to facilitate the prediction of potential consequences and side-effects of an action, still the coherence and consistency of his actions requires him to additionally consider the ramifications of a future action for some of his other actions and intentions – both past and future ones. Many dimensions are involved in the space of options that open up in that case, even more when the agent should integrate in his distal intention the potential roles of other agents, their intentions and actions.[445]

---

[444] As an aside, the word 'drama' is derived from the Greek verb 'draomai' which means not so much to perform or play, but rather to act or to do. Not surprisingly, Greek tragedians were well aware of the relevance of their dramatic work for their audience and its engagement with actions, as Snell argues (Snell 1928). An obvious source for that observation is Aristotle's Poetics in which it is argued that author, performer and observer are all participating in the mimetic experience which a tragedy presents.

With these dimensions involved there are also many and diverging values and norms involved, making it quite hard for the agent to rehearse such complexity in his imagination and also hard to conduct an evaluation of the extremely heterogeneous consequences – including future regret. Moreover, since distal intentions are temporally extended they will probably stretch over situations, experiences and consequences that bring highly divergent results which are hard to predict: an early intermediate stage of an action might be exhausting or exhiliariting while a later stage might be rather fulfilling, or despairing.[446] It is therefore important for an agent when engaging in such imaginative action rehearsal not to prematurely abort it – although it is obviously also impossible for him to be complete and comprehensive in such rehearsal.

Such desiderata are not necessarily being met when an agent is forming intentions, plans or policies in a straightforward way. Instead, for these an agent may fall back on a limited exercise in imagined rehearsal that has learnt him how a particular course of action brings about a future that concurs with a particular self-governing policy and its consequences. Indeed, it has again been Aristotle, who has long ago pointed out how such imagined rehearsal may help increase the prudence in agents' decision making with regard to their action. Particularly in his Poetics, he points out that by drawing together many different ingredients of an action in a mythos or plot, writing or observing a tragedy in fact enables an audience to engage in an experience as if they are acting themselves: "the plot is the imitation of action" (Poet. 1450 a 3).[447] Particularly in more recent times, this idea that the plot of a narrative can be considered as an imitation or representation of action has gained wide currency.[448] As a consequence, the analysis of narrative has since been included in philosophical discourses about agency and identity.[449]

---

[445] The involvement of other agents underscores that coherence and consistency have a limited value for an agent as meta-norms only: various agents may want him to interact consistently with them while complying with the specific values they are each individually attached to.

[446] Velleman points to the fact that the emotional aspect of narrative explanation is important in that it brings the temporality of actions to the fore. He acknowledges  that Aristotle's analysis of  poetic mythos already includes the 'emotional cadence' it brings about in the audience (Velleman 2003).

[447] Bywater translates 'mythos' somewhat anachronistically as 'representation', which does not capture the act of imitation ('mimesis') in which author, actor and spectator all participate, cf. (Aristotle 1984 2320). A nice, less traditional yet adequate rendering of mimesis that appears to us to capture Aristotle's intentions is 'enactive imitation', which contributes to the rich understanding and wisdom of the audience (Halliwell 1987 Ch. 4).

[448] The relatively recent reappropriation of narrative and narrative emplotment as an important phenomenon for philosophical analysis is remarkable as it takes place in both continental and Anglo-Saxon philosophy. The introduction to the Cambridge companion to narrative implicitly offers an explanation for this resurging interest, opposing narrative to science: "Narrative, in other words, is a basic human strategy for coming to terms with time, process, and change – a strategy that contrasts with, but is in no way inferior to, "scientific" modes of explanation that characterize phenomena as instances of general covering laws" (Herman 2007 3). Dissatisfaction with science, it seems, could lurk behind the interest in narrative.

Indeed, in the present context, narrative is taken similar to intentions, namely as a representation in which actions are put central – actions in a plural sense, since narratives are generally not restricted to a single (agent's) action.[450] Yet more than when forming a particular intention, an agent who is forming a narrative's plot is establishing the 'synthesis of heterogeneous elements': "by means of the plot, goals, causes, and chance are brought together within the temporal unity of a whole and complete action" (Ricoeur 1984 ix).[451] This synthesis consists first of elements that intentions share – though in a much more limited sense – with narrative, namely 'events or incidents' that are taken together in the narrative. Second, narrative integrates also elements that do not figure in intentions – or only in a limitative sense: it contains not just the intended, planned and anticipated elements of actions or events but also those that were not intended, planned, anticipated. Finally, given these heterogeneous elements, the synthesis established by a narrative is brought about by providing a configuration of these elements instead of a mere succession. This configuration is temporal in nature and this temporality of narrative is essential for the synthesis or unity that narrative can establish (Ricoeur 1991a).[452] Taken all together, these ingredients and their configuration yield narrative's meaningfulness.

With this synthesis of heterogeneous elements narrative is providing the resources that are required for the dramatic rehearsal in imagination of alternative action intentions.[453] Given their distal nature, such rehearsal is important when an agent is

---

[449] An early review of the interaction between the philosophy of action and the theory of narrative is given in (Van Dijk 1976), focusing particularly on the discourse structure of narrative and how action components are represented. Ricoeur defends systematically the general comparability of action and text in his (Ricœur 1971), which we have brought together with Gadamer's position in (Keestra 2008).

[450] In her introduction to an edited volume on 'Practical identity and narrative agency' the author writes: "The central claim of narrative approaches to identity or agency is that the lives of persons cannot be thought of as a series of discrete, disconnected experiences or events. Rather, to be a person is to exercise narrative capacities for self-interpretation that unify our lives over time" (Mackenzie 2008 11). The narrative approach to agency as applied by us is meant as an extension of the planning approach to agency that we've analyzed above. Self-interpretation is here taken to include a reflexive 'imagined rehearsal' and interpretation of one's agency, which is required for a temporally extended and planning agent.

[451] In his conversations with Changeux, Ricoeur has insisted on a distinction between phenomenology and the cognitive sciences, asking "can intentions in fact be naturalized?" (Changeux and Ricoeur 2000 67). Later on he elaborates on the complexity and dynamics that intentionality entails, including a form of selfreflection. Neuroscientist Changeux suggests to study these questions by way of discussing "plausible models of selfregulation; to discuss testing the internal consistency of plans of action" (Changeux and Ricoeur 2000 69). Ricoeur then expresses "wonder about our understanding" of the relation between such models and our actual experience, yet does not deny their possibility in a principled manner. Our account here is aimed to contribute such a 'plausible model of selfregulation' in which intentionality and neural processes are integrated. We do so, i.a. by providing a neural simulation account of the role of narrative in agency.

[452] Indeed, a crucial aspect of narrative is how it allows 'games with time', in Ricoeur's words: it can put mere succession into an intentional order or reconfigure events, it can foreground some actions and push others in the background, it can change the tempo of events, etcetera (Ricoeur 1985).

forming distal intentions as he should try to establish a comprehensive representation of an action's effects and ramifications. Although this may appear as an impossible task, it is again in this context of narrative that a hierarchical structure can be distinguished which eases it. However, in comparison to our previous discussions, this narrative configuration of actions is only partly overlapping with the intentional cascade which we've established so far.

The narrative configuration of actions does extend this hierarchy of intentions as it invites the agent to reflect upon his agency at higher hierarchical levels than discussed above. For the hierarchical levels that Ricoeur distinguishes are the level of 'practices', the level of of 'life plans' and the level of the 'narrative unity of life' (Ricoeur 1992 153-163). Particular the latter is requiring a type of synthesis of heterogeneous elements that is not included in the account that did range from motor intentions via proximal to distal intentions. The consistency and coherence that the intentional cascade has been shown to support does require the agent to consider his intentions both temporally and contextually broader than we described so far. Let us consider the less wider levels before elaborating the widest narrative level, the level of the unity of an agent's life.

Were distal intentions so far limited to plans or policies, 'practice' in Ricoeur's sense refers to a broader unit of action as it includes reference to a rule governed, general set of plans and policies that belong to a certain practice, like the practice of music making or teaching. Although simple actions are comprised in such a practice, practice itself is a 'global action' consisting of a 'nesting relation' between many subordinated actions (Ricoeur 1992 153-154). Important to note is that, apart from its greater complexity, a practice is constituted by intersubjective rules that have become internalized – often through interactions - by an individual agent, although an agent might be deviating from or neglecting such a rule.[454]

Taking these characteristics into account, we can expect tradition to play an important coordinating and organizing role in such practices. In fact, an action is always navigating between tradition and innovation: tradition provides constitutive rules and norms for a practice which are inevitably always caught up in change

---

[453] The imagined actions contained in narrative are still 'imitations' of an action in this account, and thus still "subjected to the constraint of the corporeal and terrestrial condition" (Ricoeur 1992 150) – they can be integrated in an agents sculpted space of actions and are not mere unbounded phantasies.

[454] In his psychological account of narrative as a form of mental organization used by all humans, Bruner counts among the properties of narrative its particularity, composability, diachronicity, canonicity and breach, and its normativeness. Canonicity and normativeness are related in that these also contribute to narrative's understandability, even if a narrator can allow himself to deviate somewhat from such norms (Bruner 1991). Comparable is the 'narrative practice hypothesis' which aims to offer an explanation for the development of children's capabilities to make sense of other people's actions by understanding and mastering the narratives that accompany and elucidate these actions (Hutto 2007 ; Gallagher and Hutto 2008).

and modification. [455] Such a practice is generally rather implicit, which renders it a 'prenarrative quality' or a 'narrative prefiguration' (Ricoeur 1992). In other words, the representation involved in a narrative account of a practice articulates an implicit and complex hierarchy of component actions and it the result of both a sedimented tradition and a parallel history of inevitable innovations (Ricoeur 1991c).

The second hierarchical level in this account of action is even more comprehensive than a practice is, consisting even more of a configuration of heterogeneous components. The 'vast practical unit' that is at stake at this level is the 'life plan', which consists of distal intentions with which an agent shapes his 'professional life, family life, leisure time, and so forth.' Here we find more explicitly an important role for the kind of imagined rehearsal of different lines of actions, in which action plans are included as mere components, for it includes: "the weighing of advantages and disadvantages of the choice of a particular life plan on the level of practices"(Ricoeur 1992 158). What is emphasized in this context is that such life plans are not only determined by the relatively stable and predictable practices which shape an agent's actions, but also by the much less stable, predictable and representable 'unity of a life' – the third level in this account.[456]

Are practices and life plans already rather difficult to determine, requiring an agent to develop representations that are complex, dynamic and consisting of heterogeneous component actions, isn't it impossible to configure a 'narrative unity of a life'? Given alone the involvement of many situational and intersubjective factors, isn't the unpredictability of future trajectories of him and his actions such that it makes such an effort meaningless? Nonetheless, given the expected modulating effects of developing such high level representations on the lower hierarchical levels of an agent's intentional cascade, narrative probably still contributes to the development of an adequate sculpted space of actions. That is, by engaging with a narrative account, the agent is required to develop configurations that are more or less successful in accounting for his past actions and influencing his potential future actions, irrespective of his principal finitude and the corresponding fragility of these distal intentions and actions.[457] In sum, we expect

---

[455] This inevitability of modification and change is connected to the ontological fact that an action has always an autonomy and objectivity comparable to those of a text, which is inevitably interpreted independently from the author's subjective intention (Ricoeur 1971).

[456] Similar to Bratman's methodological priority of future-directed intentions (Bratman 1987), in Ricoeur's account we find a primordial role for the configuration of the unity of an agent's life. Coherence relies to a large extent on the unity over longer periods of time in both accounts.

[457] Discussing historical narratives, Ricoeur elsewhere considers how (re-)configurations of the past are intimately related to configurations of future actions and events, and conversely, or "the complex interplay of significations that takes place between our expectations directed toward the future and our interpretations oriented toward the past" (Ricoeur 1988 208). It is also emphasized how narratives – also private ones – are never closed for "we can see how the story of a life comes to be constituted through a series of rectifications applied to previous narratives" (Ricoeur 1988 247).

precisely by his articulating – or rather: attempting to repeatedly articulate modified versions – a comprehensive account of his actions at several levels of specificity that he can aim to consistently sculpt his space of actions. If not for its correctness, it is such a contribution that supports the demand for a "narrative unity of a life, under the sign of narratives that teach us how to articulate narratively retrospection and prospection" (Ricoeur 1992 163).

Important to note is that building such a narrative is not a task that an agent only needs to do once and for all, or that after having at a certain moment presented a 'narrative unity' of his life he only needs to add novel events and actions to it. On the contrary, an agent will need to revisit his narrative self-account as he will discover that a particular configuration of events, intentions and actions has turned out to be relevant for his life, even though it wasn't part of his narrative before. In narrating and renarrating his life's events and actions, he continues to reconfigure the intentional, temporal and normative elements that constitute his identity – which he may do with some freedom, although coherence and consistency between his narratives should be guarded, as well.[458]

As such, engaging with narrative, even if it should recount something as overwhelming as his 'unity of life' is at the same time a way for the agent to get to know and to interpret himself and in so doing results in a 'refiguration of the everyday concrete self' (Ricoeur 1991d). The term mimesis, which plays an important role in the account of narrative's contribution to the agent's identity, captures this double role of both representation or imitation of his life and of its arrangement with the help of the imagination (Ricoeur 1991c). Narrative, we can conclude, is not an addition to an agent's agency or intentional cascade without any impact on these but should instead be considered an integral component of it.[459] Indeed, narrative offers contributions to the sculpted space of actions in addition to those that we have considered earlier. We will characterize those in the next section.

---

[458] Narrative, according to Ricoeur, can be considered as a 'second-order discourse', referring not directly to the events and actions of an agent's life. Although it does to some extent represent events and actions, it's second-order nature allows some liberty in their configurations, upon which the agent's identity also builds (Gregor 2005).

[459] Although the role of narrative in a subject's life is often taken to support self-understanding and reflection, in the current context its role is understood as supporting the consistency of his actions. These two roles can enhance each other, as is suggested by Schechtman in writing: "The basic features of the hermeneutical view are most easily explicated by looking at the strong connections it draws between selfhood, narrative, and agency. Selves are fundamentally agents on this view, and agency requires narrative" (Schechtman 2011 3). However, with regard to the critique of the assumption that a subject's selfhood would require it to employ narrative (Strawson 2004), it is relevant to emphasize how here our primary interest is in narrative's role of supporting agency. Indeed, we concur that it is advisable not 'to expect too much' from narrative particularly with regard to the construction of personal identity (Lamarque 2004).

### *4.1.5  Narrative as simulation of distal intention, contributing to a sculpted space of actions*

The coherent yet fast and flexible performance of our expert singer is supported by the establishment of a cascade of representations of actions based upon his previous experiences. Be they in the form of motor intentions that yield prepotentiated motor responses to particular affordances that a given stage scene present to him, or the distal intention to avoid smokey environments during a performance period, they sculpt his space of actions at several levels of specificity. Nonetheless, his experiences are necessarily limited and may not be adequate in preparing him to act appropriately in all future situations – not only for those future events that are unpredicted but also for those future situations that he anticipates or expects to happen. Given the complexity of each situation, offering multiple lines of action, it is a challenge for our singer to imagine what courses of action would in a future situation be open for him and how these would unroll, both in light of his past actions as in light of their potential consequences. What resources does engagement with narrative provide to him in these respects, how does "narration serve as a natural transition between description and prescription" (Ricoeur 1992 170)?

To begin with, when narrating an action, an agent is not just recounting all he can remember concerning it. Instead, it is a simulation of an agent and his experiences embedded in a context, a simulation which is always dependent upon some form of selection and abstraction as only a limited amount of - relevant - elements are represented in it. Another important aspect of narrative is, as mentioned above, that it configures events, actions and other relevant elements in such a way that a meaningful synthesis in terms of actions is created instead of a mere succession of events (Ricoeur 1988). One may wonder what use such a simulation can have for deciding about future action options, if it is so much dependent upon processes like the abstraction, selection and configuration of elements. What can such simulation bear upon the agent's every formation of action intentions?

However, precisely because narrative simulation does require an agent to constrain and synthesize the elements that go into a narrative about himself and his actions, it offers him resources for 'explanation by emplotment' (Ricoeur 1984 181). Important with regard to this process of giving a narrative account of his actions is the fact that in doing so his identity is always both challenged and reaffirmed. Is the mere development of such a simulation in itself already a sign of the agent's capability of both distantiation from and reappropriation of himself, the narrative does enforce upon him a certain identification with his actions. This identification has two aspects.

First, we assume the identity of an agent who is the subject of a narrative to have

a certain constancy, to remain the same. Without him remaining the same, any form of coordination and organization of actions, or care for means-end coherence and avoidance of counterproductive actions would be impossible. However, this refers only to a limited form of identity which is that of 'sameness'. Sameness refers primarily to the agent's corporeal constancy: "bodies are indeed eminently identifiable and reidentifiable as being the same" (Ricoeur 1992 33). Sameness, applied to an agent who is the subject of a narrative account of his actions, does emphasize his permanence in time which is a precondition for the other aspect of identity, about which we come to speak in a moment. Before doing that, it is relevant to note that this self-constancy is fundamental not just for the individual agent but also for his interactions with others over time and the coordination and organization these require. This is the case even though narrative can challenge an agent by presenting him alternative configurations or action options, tempting him to deny his sameness. Although he might be used to making promises of future actions, he might well also ask the promisee: "Who am I, so inconstant, that *notwithstanding* you count on me?" (Ricoeur 1992 168, italics in original).[460] This brings us to the second aspect of identification, which refers to the agent's self-hood rather than sameness.

Notwithstanding the corporeal constancy, which supports the sameness of an agent over time, agents are also notoriously inconstant. Ricoeur uses the example of a promise which an agent can make, allowing him to create a relation between past and future actions which is at the same time, however, vulnerable to the agent breaking his promise and this relation. Sameness and selfhood are put in what Ricoeur refers to as a dialectical relation, in which selfhood presupposes sameness while at the same time extending this agential constancy with a capacity for diversity and deviation. Even though the agent's longstanding attitudes and policies - those which represent in Aristotelian terms his character[461] - constitute his contancy and sameness,

---

[460] Ricoeur offers a critique of the Lockean account of identity by arguing that we need to apply a more dynamical notion of identity – as is presented by his narrative account of identity – than Locke has presented (cf. (Ricoeur 1992 143)). Bratman in turn does provide an analysis of agential identity in terms of 'Lockean continuities and connections', which Locke had originally grounded in the agent's memory and Bratman bases upon the self-governing policies that an agent has adopted (Bratman 2006b 29, 42, 59, 82, 270). The specific challenges to that identity that the narrative account provides notes are indeed absent from such a Lockean account, leaving agential identity less dynamic.

[461] Ricoeur refers to character as 'the set of acquired dispositions and sedimented identifications-with' (Ricoeur 1992 167). Character as an essential ingredient for moral agency has been introduced by Aristotle, who has coined the notion of 'hètos' in his ethics. With its long-term dispositions and attitudes, it is the source of constancy in an agent's actions and of his emotional responses afterwards. It has evolved over time, partly due to the voluntary choices that the agent has made (Ethica Nicomacheia III, 5). Nonetheless, Aristotle acknowledges that there are also constraints at work, some of which of natural and others of social nature (cf. (Leunissen 2012)). Sherman has pointed out how we can distinguish 'character coherence' and 'temporal coherence' in Aristotle's ethics, of which the former not just refers to rational coordination and organization of actions but rather to the moral unity of a life (Sherman 1989). This is quite similar to the two forms of identity that Ricoeur refers to (Ricoeur 1992).

by employing our narrative imagination we can envision how this selfhood is shown not to be equal under all conditions, indeed: "This mediating function performed by the narrative identity of the character between the poles of sameness and selfhood is attested to primarily by the imaginative variations to which the narrative submits this identity" (Ricoeur 1992 148).

According to this account of narrative identity, an agent cannot choose to avoid such variations with respect to his identity if he develops such a narrative about himself.[462] Did we note above that Bratman did leave an extra role for the "dramatic rehearsal (in imagination) of various competing possible lines of action"" (Bratman 2006b 150), narrative identity is for Ricoeur not an extra option on top of sameness or selfhood that an agent can choose to engage with, or not. The narrative unity of life, resulting from his attempt at configuring a narrative emplotment of many heterogeneous elements that is never complete and homogeneous, contributes essentially to the agent's identity and functioning as an intentional agent.[463] By developing a narrative unity of his life, with its many past and future actions and interactions and their diverse ingredients, the agent will always be affected by this form of self-understanding as it results inevitabley in a "reconfiguration of life by narrative" (Ricoeur 1991a 26). According to this account, narrative is more than just another tool for developing and evaluating distal intentions: when he is engaging in narrative in order to simulate his distal intentions in a comprehensive sense, he is at the same time reconfiguring his identity in a way that is not included in Bratman's account of distal intentions and policies.

Moreover, this reconfiguration of the agent's identity takes place in the wake of a narrative process that employs configurational schemas which are not completely developed de novo by him.[464] On the contrary, as was already noted in the previous

---

[462] Not all agents will have an equal interest in developing such self-narratives nor do they all experience a similar need to do so. Galen Strawson even suggests that narrative is more an 'affliction' than a 'prerequisite' for a good life (Strawson 2004). Agreeing with that remark only to a small extent, we would rather suggest that perhaps those lives that are lived rather according to traditional schemas and in stable intersubjective relations require less engagement with narrative than those lives of which the complexity and dynamics increase the risk of a lack of coordination and organization between actions.

[463] Carr also points out that as much as narratives are based upon selection and selective attention, so are subjects also trying to govern their lives by selectively attending to and planning for some over other elements and actions – even if they will never completely succeed in avoiding unselected events etc. (Carr 1991).

[464] An agent will continually engage in such narrative recounting of his life – with its 'elusive character' - and actions as a way to "organize life retrospectively" for which the intersubjectively and historically developed schemas are employed, as he is: "prepared to take as provisional and open to revision any figure of emplotment borrowed from fiction or from history" (Ricoeur 1992 162). Given the fact that his life is always engaged in intersubjective interactions it is plausible that such traditional schemas or configurations will fit coherently to his life as an agent. Obviously, the 'models' that are handed down by tradition are not withstanding innovation as they can be taken to "provide a guide for later experimentation in the narrative domain" (Ricoeur 1991a 25), as in the domain of action according to our account.

section, the schemas are handed over by tradition. As a result, when the agent is implicitly reconfiguring his identity by developing a narrative account, he is simultaneously both sedimenting and innovating a traditional schema upon which any such narrative about an agent and his actions is based.[465] The same dialectic between constancy and inconstancy that we have observed with regard to the agent's identity therefore applies to the schemas that he always uses for the articulation and explanation of it.

To resume these last sections: we have observed how an optimal coordination and organiation of his actions seems to force upon the agent a process of simulation of ever higher levels of distal intentions and even a narrative self-account. Doing so is inevitably a challenge to his identity in that such a narrative self-account also results in some reconfiguration of it, as much as it also affects the traditional schema that is involved in such an account.[466] Relying on his sculpting the space of his actions when planning distal actions, therefore, the agent both employs and reconfirms its pre-existing characteristics while modifying these at the same time. As speculative as this may sound, it is intriguing to find that empirical evidence regarding the prevalence of simulative processes in humans does largely support this process description, based here primarily upon philosophical analysis.

## 4.2  Distal intentions and narratives and their reciprocal interactions with cognitive mechanisms

Our investigations so far have demonstrated how our expert singer can rely upon the establishment of a sculpted space of actions when he is rehearsing and performing one of his operatic roles. Different types of kludges will develop during the many hours of practicing, rehearsing, and reflecting upon these roles and play a major role in his expertise, like when he must prepare a new performance.

When preparing a new performance of Don Giovanni under a new director and with a new set, our expert singer can mentally rehearse his part by developing simulations of it while employing his memories and experiences so far and including his established intentions, preferences and choices. Checking for coherence in his

---

[465] There is a mutual dependency of actions and emplotment: human actions are in a crucial sense mediated by narrative plots and narrative is – as Aristotle already observed – essentially about actions. Ricoeur writes about the human action that if it "can be recounted and poeticized, in other words, it is due to the fact that it is always articulated by signs, rules, and norms" (Ricoeur 1991c 141). In the previous section, section III.4.1.4 we've referred to the different levels of action – practices, life plans, unity of life – that are also determined by a socio-cultural context. Ricoeur even contends that "[t]raditions are essentially narratives" (Ricoeur 1988 260)

[466] As Scott-Baumann argues, Ricoeur systematically navigates between what is called the "totalizing tendency of a metanarrative like Hegel's" and the "total rejection of such narratives by Lyotard", which applies also to the narrative identity of an agent (Scott-Baumann 2009 92).

actions and singing, he will develop a representation that is hierarchically structured, organized with acts, scenes and concrete actions, which are also temporally organized. This relatively effortless simulation may come abruptly to a halt when he reads in the stage directions that he is to violently attack Zerlina and walk around half-naked. These directions disturb his simulation when he realizes that they conflict with some of his other intentions or long-standing attitudes that he has developed and probably also publicly expressed. Instead of simply continuing the simulation of his performance, he is now forced to consider and articulate whether compliance with these directions would undermine his personal or artistic integrity, for example. Doing so requires him to find out what eventually happens to both the Don and the maiden during the play – whether some poetic justice is being done, or not - , if the piece is broadcast or otherwise can be seen by young children, whether his physical appearance forbids its nakedness, whether him doing this performance would contradict previous interview statements of his, and so on. Obviously, this is no longer a matter of further elaborating a particular action representation, as it requires him to analyse many of his intentions and multiple previously performed actions, to consider his preferences and norms, to simulate potential interpretations of his performance by an audience and then to check for coherence and conflicts between these ingredients.

It is in such a case that narrative simulation will play a crucial role and spell out the necessary web of intentions, draw upon relevant past histories and look forward to potential future actions and events. Such a narrative simulation is a rather comprehensive and time-consuming task that is never really completed. Given our interest in kludge formation as a result of learning and development and in the role of kludges in facilitating an expert in making fast and flexible responses, we are now facing the question whether the processes and representations involved in such narrative simulation can undergo a similar process of kludge formation, or not? Before answering that question, remember that we've highlighted three – of the original seven - kludge characteristics in the present Part. First, we noted that kludge formation is observable in the performance properties of a particular function, as performance usually becomes more coherent and consistent. Nonetheless, the kludge's implementation in terms of represented information or of neural correlates cannot be derived from these properties. Second, a kludge is established usually by the re-use or recruitment of already available component mechanisms and it can itself subsequently be employed as a component in a functional mechanism. Indeed, when a kludge is further employed it becomes ever more generatively entrenched in the dynamical system or organism. Finally we noted that environmental and even cultural information can become integrated in a kludge's performance – like conceptual

information or a ritual object that automatically modulates expert action.

These kludge characteristics were found in chapter III.2.2 to apply to the development of sensori-motor representations in the form of templates with free slots underlying an agent's motor intentions, enabling experts to perform habitual actions both more efficient and more flexibly than novices (Gobet and Simon 1996). Well established proximal intentions, we later found in chapter III.3 ff., involve more elaborate hierarchical representations – like Structured Event Complexes (Grafman 2003) - that are employed and modulated by several cognitive processes, like the automatic contention scheduling process and supervisory attention (Cooper 2002). Supervisory attention or other processes can also modify or even generate new representations that might eventually be integrated in newly established kludges.

Indeed, in II.4 we've discussed how a particular 'simulator' can be considered to be an example of the kludges that our brain appears to be capable of establishing. For our discussion of simulation we did rely upon Barsalou's theory of perceptual symbol systems (Barsalou 1999c) and later elaborations of that simulation account. In that context, we noted two features that are relevant to recall here. This simulation account argued that sensori-motor resprensations are not stored as a whole, but that features of those representations are stored in a highly distributed manner across the brain. This process of coding, storing and reconsolidating memory features can in turn be modulated by other processes, like by the attention an agent pays to one or more of those features (Barsalou 1999c ; Barsalou 2009). Second, agents can activate or reactivate through multiple cognitive processes a complex of many different features that are more or less associated with each other and in so doing simulate past, future or imaginative experiences: "simulation is the re-enactment of perceptual, motor and introspective states acquired during experience with the world, body and mind" (Barsalou 2009 1281; cf. Barsalou 2008).

In these last sections of our final Part, we will consider how the development and consideration of distal intentions relies upon such simulation processes. Such simulation processes should then underly not just the articulation of a complex and temporally extended action intention, but also the imaginative rehearsal or narrative of a multiplicity of such intentions and integrating contents from the agent's past and visions of the future, too. In this way, simulation should facilitate an agent to anchor and specify his intentions not just in a particular situation *hic et nunc* but also in intersubjective situations in a more distant future. Of particular interest to us is whether and how such simulation contributes to the agent's overall sculpted space of actions and is eventually observable in his actions, as well. This brings us back to the issue of kludge formation, even if by now we may expect that when this is happening

with respect to narrative simulation, too, its implementation to be more complex as well.

Indeed, in these last sections we will present some relevant evidence confirming the assumption that kludge formation also applies to this level of distal intentions and narrative simulation, contributing again to the coherence and flexibility of an agent's actions. It may as well not be surprising to find that the kludge finally appearing at this level is a rather complex neural network. Indeed, this so-called 'default mode network' has been discovered only recently (Raichle, MacLeod et al. 2001) and has been meanwhile found to be directly or indirectly involved in many different cognitive processes, which explains why it is still a matter of debate what its components and properties are, and what not. Notwithstanding such differences, there is agreement that its "intrinsic activity instantiates the maintenance of information for interpreting, responding to and even predicting environmental demands" (Raichle and Snyder 2007 1087). In doing so, this default mode network (DMN) constitutes for the agent a kludge that is relatively stable and enables his acting coherently and flexibly without always demanding explicit reflection. Concurrent with our reasoning, the DMN is determined both by the lower levels of the intentional cascade as by the level of distal intentions, like when it is modulated by the agent's explicit narrative simulation.

Let us take a look at some empirical evidence for the contribution of our simulation capabilities to our performances as a planning agent, by first paying attention to the simulation of a single distal intention and then the narrative simulation of a more complex web of intentions.

### 4.2.1  From memorized experiences to the simulation of future actions

As mentioned above, experiences help an agent to learn performing certain motor skills or more complex behaviors in a flexible way. The sensori-motor representations or more complex Structured Event Complexes that are the result of expertise and learning enable an expert to perform differently in comparison to novices. These insights offer a picture of memory or stored information different from traditional accounts of memory, wich generally contended that memory was meant to keep an archive of the past.[467] In contrast to this, more recent accounts emphasize the role of

---

[467] Indeed, as early as 1932 Bartlett argued that the function of memory should not be considered to archive and reproduce a faithful picture of the past (Bartlett 1995 [1932]). Research has confirmed that there are many specific vulnerabilities or modulating factors involved in the three phases of memory, i.e. the encoding, consolidation and retrieval phase (Straube 2012). Instead of faithfully reproducing the past, evidence shows the reliability of memories to correlate with the likelihood that they are needed for future tasks, implying that "the memory system tries to make available those memories that are most likely to be useful" (Anderson and Schooler 1991 400).

memorized information to enable future action.[468] Indeed, it is argued that memory must have an adaptive value and therefore help an organism to better prepare for future actions (Klein, Cosmides et al. 2002). Arguments to that effect have not only relied upon insights about biological evolution. More generally – which is discussed particularly in Part II - complex and dynamical systems must be capable of letting environmental information determine partly its development in order to enhance the speed and flexibility of its interaction with its environment (Simon 1973). A very simple example to which we referred earlier is the relatively rigid mechanism of imprinting in goose chicks, which has been shown to be sensitive to contingencies of the environment, determining the drastically different trajectories of those chicks' lifes that were imprinted by a human instead of an avian being (Wimsatt 1986). Such animals are not capable of explicating what information it is that they have stored or how they employ and affect their behavior accordingly, but these limitations do not stand in the way of some form of adaptivity via acquired traits that are based upon represented information.[469]

Also on the neural implementation level, there have associations been found between the processes involved in memory and in the simulation of future actions. At a more fundamental level, both long term potentiation and the principles of Hebbian learning apply generally to those processes and indeed can explain why experience based learning is associated with specific dispositions that determine future cognitive and behavioral responses (Kandel 2009). At a more complex level, similarities in the neural networks between memory and preparation for the future have been found as well. After decades of research, different taxonomies of memory haven been proposed that correspond not only with different neural networks but also with properties regarding the simulation of future actions (Nadel 1992 ; Sherry and Schacter 1987).

---

[468] Obviously, this is only one observation underlining the theory that most, if not all, cognitive functions have evolved such that they are preparing the organism for future actions. Several theories have been presented, which reinterpret the evidence about cognitive functions and brain processes along those lines, like the theory of predictive coding (Friston 2005 ; Clark 2012), the theory of the proactive brain(Bar 2009 ; Pezzulo and Ognibene 2012), or theories assuming the prevalence of Bayesian processing in the brain (Barsalou 2011 ; Colombo and Seriès 2012 ; Kording and Wolpert 2004). There is some overlap between these theories, as Bayesian processing allows the brain to operate as a "prediction machine' (Clark 2012).

[469] The notions of information, representation, and cognition implied by this example are all debatable – see e.g. (Bechtel 2008 ; Dretske 2003 ; Gärdenfors 2004a ; Keijzer 2002 ; Piccinini and Scarantino 2010 ; Rowlands 2012). Engaging with that debate falls beyond the scope of this book. However, we consider the fact that under certain conditions the goose chicks are responsive to a set of features, related to each other in a particular configuration, to be a matter of their being responsive to a representation of natural – non-conventional - information. The perception and cognitive processing of these features and their configuration allow for several mishaps, leading to dysfunctional responses, as when a chick would start to escape instead of approach the animal which appears to be its parent.

Indeed, the 'neural machinery' that is recruited for remembering the past turned out to be largely overlapping with the machinery involved in the simulation of the future (Schacter, Addis et al. 2007). An interesting example of such a correspondence has been observed when an agent aims to influence his future action with counterfactual thought, as when he simulates how a particular action might have been different with regard to action properties or outcomes. Overlapping neural network activations have been interpreted as involved in representing the Structured Event Complex pertaining to the specific action (Barbey, Krueger et al. 2009).[470] In sum, simulation of one's future action is effective, it can be said, by developing or modifying an action representation that is then stored as a memory and offers a preparation for its future implementation in motor behavior – thus underscoring the interdependence of memory and simulation of the future (Papies, Aarts et al. 2009).

Such insights about the general association of a complex and dynamic system's capability to store representations and its preparation for the future did not preclude other perspectives on intentional action planning. Indeed, it was initially thought that intentional action planning in humans depends upon specific capacities that were held to be exclusively human – in particular human language and episodic memory, which both are responsible for specific representational functions. Meanwhile, various lines of evidence showed otherwise and forced the nuancing of this assumption. For example, evidence from amnesic patients and other subjects have shown that planning for personal future actions can survive in the absence of episodic memory, demonstrating that relevant information can also be represented in semantic memory and still be involved in simulation processes and be adequate for some forms of planning (Klein 2013). Similarly, evidence in animals like scrub-jays and primates has suggested that even they are capable of some form of 'foresight', of planning future actions (Suddendorf and Corballis 2007).[471] So notwithstanding apparent differences between animals and humans with regard to the possession of language, to metacognitive abilities and to the available types of memory, they do appear to share some capabilities for action

---

[470] There are psychological differences between rembering the past and such simulations, to be sure. The clarity and vividness of remembering the past to be stronger and emotional intensity does not appear to be equal for all conditions (De Brigard and Giovanello 2012).

[471] A remarkable observation was that of a male chimpanzee in a zoo that collected during quiet mornings ammunition as it had the 'foresight' that he would need this when in the afternoon he would be irritated by human spectators shouting and grimacing at him. Remarkable was that the observed motivational state of the chimp during collecting was different from the state he later would be in when using the stones, whereas most planning in primates is dependent upon the similarity of their motivational states at both the planning and execution phase of an action (Osvath 2009). This observation notwithstanding, it is widely held that such future-directed planning on behalf of anticipated, future motivational states is not available to primates (cf. Bischof (Bischof-Köhler 1985 ; Pacherie and Haggard 2010).

planning involving the representation of such action.[472]

Observations in children and in novices confirm that previous experiences at first only implicitly contribute to their improving performances, with explicitation being left to later stages. For example, in the beginning of their motor skill learning, they are incapable of articulating the relevant representations involved or of consciously governing their use in future actions.[473] However, according to the neuroconstructive account of skill learning - which has been discussed in chapter II.2 ff., - the proceduralization of a skill is followed by its explicitation: after automatizing the skill, its explicitation eventually allows a skilled child or expert to make connections between different representations associated with the skill, to draw analogies, and so on (Karmiloff-Smith 1992). This productive use of representations is enabled by the process of Representational Redescription which obtains during learning, it being associated with the cognitive and behavioral differences between novices and experts (Clark and Karmiloff-Smith 1993 ; Cleeremans 1997 ; Karmiloff-Smith 1990 ; Mareschal, Johnson et al. 2007).

These observations are perhaps not surprising in that they just demonstrate how stored information allows an agent to better prepare for future actions. Moreover, they concur with the observed parsimony with which brain mechanisms develop and perform their functions, generally by re-using or recycling previously established components and by forming kludges as a result of developing and learning, as was discussed in our previous Parts. However, it is relevant to note that action representations can have such preparatory value even if past and future actions are not identical. In between their involvement in past and future actions, such stored information can apparently be modified flexibly, yielding useful representations under changing conditions. The question that then presents itself is how are stored action representations used in the simulation of future actions, like when an agent is engaged in forming distal intentions or in narrative? Can we observe a form of kludge formation in this context, tto? Moreover, we would like to know whether these simulations are also affecting the sculpted space of the agent's actions and whether this is observable in his behavior. These questions will concern us in the next sections.

---

[472] Developing a rodent model of episodic memory, Crystal reviews evidence in rats of their capability of not only remembering specific events in the past but also of employing these representations with some flexibility (Crystal 2013).

[473] Computational and observational studies show, however, that skill learning can be facilitated by the use of both implicit and explicit learning. Explicit learning initially depends upon the articulation of representations of relevant information and tasks. Yet, the two types of learning can at times also conflict to the detriment of skill learning (Sun, Slusarz et al. 2005)

### 4.2.2 Distal intentions in the intentional cascade: from action control to mental time travel

Let us first return to the intentional cascade, which has provided the structure of our discussions so far, even if we have observed some limitations of the distal intentions as presented there. When discussing the functional role of intentions in her account, Pacherie focuses on the different forms of control that are associated with motor, proximal and distal intentions respectively. Control is implemented in the intentional cascade in the form of comparators, comparing the properties of representations of a desired state of affairs, of the action that is supposed to realize that state and of the eventual outcome of that action (Pacherie 2008). The idea lying behind this model is that multiple representations of an action are being processed simultaneously, representations that can also partly be shared by different cognitive processes. It has been Jeannerod who has worked on the theory of shared representations.

The theory of shared representations was presented in order to account for various phenomena, ranging from significant overlap in neural activation patterns during the observation and performance of similar actions to the delusions of action control that occur in schizophrenic patients (Georgieff and Jeannerod 1998). These findings were added to other experimental results that together lead researchers to conclude that even though action representations are processed during different functions – like when actions are being observed, imagined, verbalized and performed (Grèzes and Decety 2001) – the neural implementation of these representations overlap to a large extent. Indeed, as Jeannerod did conclude, the greatest difference between all those forms of 'simulation' – in his words - and the performance of a real action may just be the lack of motor activation in order to physically execute the action representation in the former (Jeannerod 2001).[474]

Instead of elaborating upon this notion of simulation, Pacherie has chosen to integrate into her intentional cascade several comparators, each of which are focusing on similarities and differences between representations of a state or action. Neuroscientific and computational studies have indeed suggested that representations are employed in parallel in preparation of and during an action and that the – implicit - comparison of such representations support the continuous monitoring and adjustment of an action (Wolpert, Ghahramani et al. 1995). In the intentional cascade model, three comparators are assumed to be part of that model. One comparator is held to compare the actual feedback that the agent receives with the initial desired state he had when

---

[474] This explains why exercising motor movements through imagery is effective in ameliorating their real performance, as is for example demonstrated in motor imagery exercises in rehabilitating stroke patients (de Vries and Mulder 2007).

developing his intention. Another one allegedly compares his motor command with the actual outcomes of the motor action and is connected to his proximal intention. A third comparator is not relying upon actual motor representations as it involves the agent's representations of a desired state and a predicted state that would result from his action. For this reason, Pacherie argues, this comparator and the distal intentions involved in its function must be made in a representational format that is not constrained to motor properties or perceived situational properties. Consequently, this third comparator is assumed to process representations in a conceptual format (Pacherie 2008).

Because of this format, distal intentions can have some influence on the other representations involved in the intentional cascade. The explanation for such an influence is that the different types of representation, expertise and sensory information are all integrated in a Bayesian framework. Consequently, all representations involved in prediction, in feedback, in comparison and other roles are always weighted, based upon previous experiences and learning processes. Distal intentions are apparently capable of influencing the weighing of components of the representations involved in proximal and motor intentions, which is an important part of their efficacy (Pacherie 2008).

The control or regulation that distal intentions exert in this way is being referred to as rational control (Pacherie 2008). Rational control is distinguished here in the dual form of 'tracking conrol and 'collateral control', as proposed by (Buekens 2001). It amounts to controlling for the intended effects of an action and not unforeseen side effects. Since this occurs at a larger time-scale than proximal and motor intentions and can involved situational properties that were not foreseen at the time of the formation of a distal intention, the representational format of the other two types of intention would be less effective. Distal intentions can accordingly contribute to the coherence between the agent's reasons, intentions and actions, as they allow a comparison between the action properties as predicted by this rational control and its eventual outcomes and enable the agent to adjust his action if necessary. (Pacherie 2008).

Such comparators can only play their important role in the intentional cascade if the representations involved are sufficiently concrete and precise for such comparisons to be performed. Especially in the case of distal intentions this assumption is questionable, as predictability of all representations involved will decrease as a function of the distance involved to the future. With the decrease of predictability, establishing concrete representations of desired and predicted states, for example, will be more difficult and as a consequence the comparison between the two makes less sense. This may be the reason that in a more recent and adapted version of the intentional cascade,

coauthored with neuroscientist Haggard, distal intentions are considered not so much in terms of rational control but as a form of mental time travel.

With this shift to mental time travel, attention is paid to the source of the contents of distal intentions as to the question how these contents are employed such that they contribute to new intentions. The mental time travel capability is defined as a combination of autonoesis and the configuration of a particular event: an agent is constructing an event in which he is himself the subject (Pacherie and Haggard 2010).[475] Explaining how such configuration is carried out, the authors refer to an influential account of mental time travel, according to which memorized events or their features are used, being stored in semantic and especially in episodic memory, as: "one further needs to be able to combine and recombine existing elements" (Suddendorf and Corballis 2007 307).[476] Distal intentions being formed in this way, differences between them obtain.

For one, distal intentions are not all equally comprehensive and detailed. Depending partly on the extent to which a distal intention relies on a full episodic memory of a corresponding action, it will specify many details of the action and environmental conditions that should trigger it, or not.[477] In the latter case, the agent leaves many details of the action – like when and how it must be performed – open to a later moment. Indeed, the authors suggest that each agent will find himself somewhere on the continuum between a 'neurotic planner' and an 'optimistic improviser', depending upon how much he likes to specify beforehand or relies upon later specification instead – with most agents probably alternating between such modes or strategies (Pacherie and Haggard 2010). With these strategies having their differential effects and their benefits and disadvantages, the authors contend that their effectivity relies on the activation of a relevant situational cue as well as a particular and desired action representation, associated with that cue.[478] Alternatively, instead of using a situational

---

[475] In their analysis of distal intentions, the authors distinguish between what-, how-, and when-decisions. For each of these they refer to some evidence that seems to be relevant for that type of decisions, yet fail to present an overarching account that can explain how those three types of decisions for distal intentions are associated, nor how distal intentions can affect – even implicitly – proximal and motor intentions. Nonetheless, they do explicitly reject the qualitative difference that Pacherie had made in her (Pacherie 2008) between the contents of distal and the other intentions (Pacherie and Haggard 2010 82).

[476] Even though foresight or imagining future events appears to rely largely on episodic memory, the requirements of foresight and episodic memory are not necessarily identical so the distinction between the two remains important (Suddendorf 2010). Besides, it has been argued that in contrast to common opinion, semantic memory can also yield such foresight, even of self-related events (Klein 2013).

[477] Developmental studies show that planning actions in the future does require mastery of action schemas or scripts, yet this mastery is not sufficient as the future will not be a mere performance of such a schema. From age 5, children were capable of truely planning an action with more detailed preparations, being more flexible to adapt to future contingencies, and so on. This development is partly explained by their increasing employment of detailed episodic memories, not just semantic memories (Atance 2008).

cue, an agent can engage in 'time-based prospective memory', aiming to anchor his future action to a particular time (Pacherie and Haggard 2010 80).[479] In any case, some automatization of such forms of planning of actions is possible.

In their account of distal intentions, the authors combine different theories and concepts. We've noticed them referring to mental time travel, which relies upon the recombination of already available representational elements. They did also refer to prospective memory, suggesting that such elements are preserved from earlier experiences and memories. Finally, they also mention in passing how an agent involved in mental time travel can imagine or 'simulate' more or less in detail the future situation (Pacherie and Haggard 2010 80). In light of our earlier treatment of simulation as a type of computation that is prevalent in many cognitive functions and our reminder of this just above, it may not come as a surprise that we've decided to continue our discussion of empirical evidence concerning distal intentions and narrative in terms of such simulation. And similarly, let us recall how we're interested in whether such simulation by a planning agent contributes to his sculpted space of actions in a more comprehensive manner than only by way of enhancing coherence and consistency between an agent's reasons, intentions and actions with regard to just a single intentional action. For in section III.4.1.2, we've argued that the agent must be able to support this by consindering a single intention in the light of his 'total web of intentions' (Bratman 1987 32). Moreover, we found that narrative simulation offers the rich resources required for such weaving of a web of intentions and its consideration in a wider context. A specific example of such a resource are the configuration schemas or models that can be used by simulation processes and which are sometimes borrowed from tradition, relieving an agent of the task of 'emplotting' or configuring his own complex distal intention completely de novo. The question is therefore what evidence there is that narrative simulation is implemented in such a way that it can indeed contribute to an agent's sculpted space of actions, like we discussed above?

---

[478] The authors observe that the effectivity of distal intentions resembles the effectivity of 'implementation intentions' (Gollwitzer and Sheeran 2006 ; Webb and Sheeran 2007), to which we referred earlier in the context of the need for anchoring and specification of a distal action via proximal intentions, in section III.3.2.1. Since the situational cue does also activate the proximal intention and since our concept of distal intention is wider than Pacherie's, we've chosen to refer to implementation intentions in that earlier context.
[479] Interestingly, the authors refer in this context of time-based action planning that a process like the unconscious action initiation found in Libet's seminal experiments (Libet 1985) is probably involved in it (Pacherie and Haggard 2010 80). It is indeed plausible to interpret Libet's findings such that the action initiation occurs as a result of an interaction between a particular distal intention and some sort of triggering cue, with Libet underestimating the efficacy of any distal intention. Cf. the suggestion made in (Roepstorff and Frith 2004) that the subjects accept the action script that has been proposed by the experimenter, adjusting their intentions to include that script and diminishing the role of their own voluntary intentions.

### *4.2.3  Simulation and the flexible imagination of a future action*

The claim that agents 'simulate' their future behavior by exploiting previously experienced and memorized events was made several decades ago. It was argued that the 'extraction' of an action program from the observation of an ongoing action is tightly associated with the construction of a 'memory of the future' when simulating a future action. Moreover, both processes recruit allegedly regions in primarily the prefrontal cortex (Ingvar 1985). Subsequently, a more comprehensive simulation theory was presented by Jeannerod, who found overlapping cognitive and behavioral properties between an agent's motor performances and his motor imagery, suggesting that both processes also largely activate the same neural pathways (Jeannerod and Frak 1999). Yet, even though Jeannerod's work plays an important part in the intentional cascade framework, Pacherie did not assign a central role to this account of simulation. Instead of focusing on simulation and the fact that representations are shared between functions, the framework assumes the presence of different representations in the cascade that are somehow compared with each other (Pacherie 2008).[480] However, building upon the arguments given in our earlier discussion of simulation as a general computation employed in the brain we will now scrutinize whether simulation can provide a parsimonous account of how an agent develops complex distal intentions and even the complex narratives that allow him to configure a complex web of events, intentions and actions while considering alternative options for future actions.

Given the present task, we will discuss a specific form of simulation that may be taken as a particular instantiation of the general computational function that simulation is held to be (Barsalou 1999c). Simulation appears to also underlie formation of distal intentions (and action plans), by employing flexibly memories of past experiences or their components, like the Structured Event Complexes discussed previously. There is some variety involved in such simulation, which correlates with the large differences that can be observed between such intentions. When comparing distal intentions, researchers have found relevant variation along several dimensions, like their complexity and detail, their likelihood, their familiarity, and so on.[481] As can be expected, depending upon the agent's previous experiences, the representation involved in a distal intention will be more or less specific along these dimensions (Schacter

---

[480]  This is somewhat strange as she did collaborate with Jeannerod, writing a paper on agency and simulation together with him (Jeannerod and Pacherie 2004). Although the phenomenological analyses of agency and ownership are important topics in (Pacherie 2008), they could have been articulated also by putting simulation central, as she has done elsewhere (Pacherie 2001).

[481]  Dimensions of distal intentions can also be associated. For example, depending upon the future task goal – climbing or photographing - and the temporal distance – sooner or later - , subjects did draw an Egyptian pyramid with less or more detail and in different sizes (Christian, Miles et al. 2013).

and Addis 2007b ; Schacter, Addis et al. 2007).[482] For example, the level of structure and detail of a distal intention appears to benefit from the amount of experience an agent has with a particular action and the context of its future performance.[483] Based upon both psychological, lesion and imaging studies, researchers have proposed that by recombining stored memory features agents are constructing a novel episode from memorized elements. Indeed, particular memory deficits in patients are often associated with particular failures in action planning, with the representations losing detail and complexity (Addis, Musicaro et al. 2010 ; Addis, Sacchetti et al. 2009). In line with the above, the 'constructive episodic simulation' hypothesis contends that memory errors should be taken to exhibit particularly the crucial function of memory with regard to future-oriented or simulation processes (Schacter and Addis 2007a). What constructive process is involved in this particular simulation?

If an agent is to respond flexibly, he should be able when simulating a future event or a distal intention to not just depend upon his capability to encode and retrieve memory features but also have the capability to elaborate potential relations between memory features for their flexible recombination.[484] Indeed do animals and humans alike demonstrate the latter capability. For example, rats were found to be capable of making transitive inference relations between two items that were included in non-overlapping pairs of items: after having been trained to prefer A over B, and B over C, and C over D, and D over E, they will choose also B over D even without training

---

[482] This finding was supported by other research, in which subjects had to engage in counterfactual thought. They had to imagine previously experienced events such that their factual outcomes were different from the imagined outcomes, the latter being better or worse, and this with large or little likelihood. Cognitive and imaging evidence shows that such counterfactual thinking largely recruits the same processes and neural networks as engaged in 'constructive episodic simulation'. Interestingly, unlikely episodic counterfactual thoughts did rely less upon the neural system associated with remembering than upon the system associated with imagining past or future episodes (De Brigard, Addis et al. 2012). Nonetheless, engaging in counterfactual episodic thought has been shown to elicit confusion or distortion of the memory for the original event, confirming that such simulations are constructive rather than reproductive processes with memory playing an important role in those (Gerlach, Dornblaser et al. 2013)Dornblaser et al. 2013.

[483] Nonetheless there is a difference between remembering the past and planning for the future, as developmental studies show. Children start only with establishing distal intentions or other forms of planning for future behavior around the age of 2,5 years. By that same age children verbally express uncertainty about future events, make use of modal terms – 'probably', 'possibly' – and show in their behavior how they're preparing for future events that are not identical to past events (Atance and O'Neill 2001).

[484] Hippocampal activity has been taken to be responsible for relational memory, which has been investigated in imaging experiments in which subjects have been asked to remember complete routes instead of mere locations, or to remember word associations, or to remember information related to a test item (Cohen, Ryan et al. 1999). Other research with adults who had to flexibly connect previously learned, yet unrelated pictures, suggests that the hippocampus is particularly involved in the configurating of novel relations (Preston, Shrager et al. 2004). This appears to be particularly the case for quick learning tasks, whereas gradually extraction of relations between memorized features relies rather on cortical activation – which explains the differences between various amnesic patients in such tasks (O'Reilly and Rudy 2001).

this relation specifically. However, rats with disturbed hippocampal connectivity will fail to infer this new relation and fail accordingly to gain extra rewards (Dusek and Eichenbaum 1997). Research with masked word pairs in humans demonstrated again hippocampal activations to be correlated with their behavioral demonstration of an implicit capability of establishing novels relations between separately presented words, without having consciously noticed and processed these words. Hippocampal activity in this task was not limited to encoding but visible during extended periods of time, including at the time of the response task (Reber, Luechinger et al. 2012). But it is not just the hippocampus that is critical for establishing such relations. Developmental studies of children's performance on remembering a past event and simulating a future event suggest an important role for the frontal lobes and executive functions in their developing capability for establishing relations between memory features and making inferences (Richmond and Pan 2013). The moment that such basic capabilities of establishing novel relations between stored memory features are in place, it can be expected that an agent's expertise with particular schemas or configurations of actions can additionally facilitate his simulation of distal intentions with similar representational structure and complexity.[485]

Not only relations between memory features matter, but also the relation to other persons, with whom interaction is often relevant for a distal intentions, matters.[486] In that case, the simulation must take into account possible responses of other persons as these might influence the action outcomes. Here again, the simulation of a distal intention has been found to cause brain activation patterns that are similar to those typically associated with viewing a situation from another persons' perspective or mentalizing (Buckner and Carroll 2007).[487] For example, it is advisable when forming a distal intentions involving another person, to take his or her personality into account, if possible. An experiment in which subjects had to imagine particular future scenes

---

[485]   In line with the evidence mentioned above concerning the role of configuring relations between memorized features, development plays a role in children's capability of simulation of distal actions as well, since this also affects their capability for relational memory (Richmond and Pan 2013).

[486]   Papineau points out that understanding other minds, whether considered in terms of theorizing or simulating, is to a large extent equal to understanding their means-ends reasoning. The evolutionary early occurrence of such means-ends reasoning, he speculates, might have been in the visual imagination of alternative versions of observed actions, rather than in language (Papineau 2006). This would be another reason for the importance of cognitively processing the hierarchical structure of actions.

[487]   The proposal by Buckner and Carroll of 'self-projection' as a central cognitive function shared by theory of mind, imagination and future-oriented thinking has received some criticism (Buckner and Carroll 2007). For example, it has been suggested that a more adequate term for this common cognitive function would be 'scene construction' as it does not assign a central role for self-related processing (Hassabis and Maguire 2007). For similar reasons, we prefer the term 'simulation' to refer to this cognitive function as it suggests an important role for previous experiences, as well.

with the participation of varying persons with different personalities – as presented through vignettes – showed that different brain regions support processing personality specific information with a central role for the mPFC (Hassabis, Spreng et al. 2013). This is yet another indication that the simulation of a distal intention poses extra demands on the cognitive and neural processes in comparison to proximal intentions.

When an agent has formed a distal intention, along the lines mentioned above, how can he prepare its eventual performance, taking into account the importance of relevant affordances involved in proximal intentions when anchoring an intention in a situation? Investigation of 'implementation intentions' demonstrate that such affordances be integrated in his distal intentions, if actions like exercising or other healthy behaviors are to be promoted. A distal intention is more effective to the extent that such a simulation specifies not just the goal of an action but also the situational conditions – when, where and how – under which it should be performed.[488] The representation that results from this is much more detailed and contains information that can serve as one or more cues for the initiation and continuation of the action (Gollwitzer 1993). Effective simulation still requires more of its agents capabilities.

Furthermore, when specifying his implementation intention, the agent should also care for its viability in the simulated future situation and its instrumentality in reaching his goal. In so doing, he can also specify potential obstacles that he might have to face, adding to the efficacy of his distal intention (Gollwitzer and Sheeran 2006).[489] In addition, several lines of research have shown that motivation matters. For example, study of a memory recollection task has shown the influence of the agent's motivation during preparatory processes. Items were better encoded and recollected when the expected reward upon recollection was higher (Gruber and Otten 2010). Similarly, even emotional responses to future situations can be regulated through implementation intentions. Such emotional self-regulation can become habitual or implicit, just like other skills do (Gyurak, Gross et al. 2011). In line with such findings concerning the relevance of motivation, research shows that implementation intentions are not capable of 'trumping' the agent's lack of motivation for a goal-directed action. Instead of programming him such that he automatically engages in such action when presented with a situational cue, his motivation with regard to the goal remains a

---

[488] Construal level theory contends that when an intention is formulated in rather abstract terms and contains less lower level details, it is perceived as pertaining to the distant future, and conversely (Liberman, Trope et al. 2007). This perception of distance may contribute to its being less efficacious than a more detailed intention, but it is plausible that the lack of detail also impedes the recognition of specific affordances in a future situation.

[489] Even though we have mentioned several reasons why distal intentions generally, implementation intentions included, rely upon speech processes, it is worth mentioning that additional mental imagery appears to enhance implementation intentions' efficacy (Knauper, Roseman et al. 2009).

prime factor in determining the efficacy of his implementation intention (Sheeran, Webb et al. 2005).

From these lines of research it can be concluded that forming an effective distal intention relies upon the formation of a relevant and detailed future situation representation and upon forging a strong association of that with a desired action (Webb and Sheeran 2007). Both component processes involved in such simulation benefit from previous expertise with the situation and with the action respectively. As much as this evidence supports our account of the relevance of a sculpted space of actions, a limitation of it is that it mostly focuses on a particular distal intention and does not address potential interactions or even conflicts between multiple distal intentions. Yet in our discussion in section III.4.1.2 we did emphasize the importance of a robust web of distal intentions with an equally stable motivational hierarchy that would enable the agent to conduct his self-governing policies such that he is not easily temped to reconsider, challenge or even deviate from these policies as this could easily be counterproductive and costly.[490] As these requirements ask for still more comprehensive cognitive processes than those involved in the simulation of a single distal intention, let us proceed to consider those.

### 4.2.4  Narrative and additional benefits of the simulation of multiple distal intentions

In the previous section we found that with regard to the simulation of a particular distal intention, we are looking at the flexible configuration of memorized action components. We noticed that such a configuration may vary along several dimensions, like its familiarity, its involving other persons, and so on. Yet we've noticed earlier that it is the heterogeneity of intentions and the corresponding action goals, the plurality of the relevant motivations and norms, and the different temporal scales on which these ingredients play their role in this web of intentions that yields more complexity. Indeed, these factors are challenging the agent's capabilities for achieving global consistency between his intentions and actions (Pacherie 2008). It may be that the mechanisms mentioned earlier may not be sufficient for achieving such results.

Developing a consistent web of intentions that can adapt to these requirements appears indeed to ask for a decisively different 'instrument' than the processes that we have considered so far, like those involved in proximal intentions and (relatively

---

[490] Baumeister observes that making distal intentions or implementation intentions also enhances an agent's success in performing his preferred actions when he is actually in a state of 'ego depletion', because such a state has an impact on conscious control yet leaving forms of automatical control intact (Baumeister, Crescioni et al. 2011).

simple) distal intentions. Proximal intentions were explained in section III.3.2.4 with a model that combined contention scheduling with supervisory processing (Norman and Shallice 1986 ; Shallice 2002). However, as we found the CS/SAS model to be limited with regard to the spontaneous generation of novel action configurations, it seems not fit to explain an agent's global consistency.

A more elaborate model was found in the theory of 'Structured Event Complexes' which leaves much more room for the free generation of novel action schemas, facilitated by the neural implementation of such SEC's in the PFC with its rich connectivity to other parts of the brain (Grafman 1995).[491] As developing a consistent web of intentions implies the inclusion of normative and motivational information and information with regard to other agents, it is relevant here to recall again an expanded version of SEC's – so-called 'event-feature-emotion- complexes -  that has been developed through studies of moral cognition and which is correlated with a wider set of neural areas involved in emotional reasoning and mentalizing (Moll, Zahn et al. 2005). Such complex representations are also available for multiple cognitive processes. Indeed, such representations can also be involved in distal intentions, relying as they do on memorized action representations, for example when 'constructive memory' enables an agent the simulation of future events or actions (Schacter and Addis 2007b ; Schacter, Addis et al. 2007). Obtaining global consistency between heterogeneous or deciding between inconsistent intentions still requires other processes and resources, however. We've offered a short presentation of narrative simulation earlier as an option. Let us now take a concise look at the cognitive processes required for this.

For such narrative simulation of his actions, an agent must be able to rely again upon his capability of yielding multiple representations – and in different formats – involved in a particular task. Such was also the result of our discussion of the process of Representational redescription, underlying several forms of learning and development (Karmiloff-Smith 1992). This representational redescription process brings along benefits for the agent, such as learning or changing domain specific rules, applying fast and specific modifications, systematic adaptation of the representational domain, and especially the"integration of activities with those of other sub-systems operating on data included in different formats" (Clark and Karmiloff-Smith 1993 492). With narrative simulation carried out in conceptual – linguistic - format, these benefits seem to be particularly pronounced, more so then when action representations would still be in the form of motor representations or in a visual format. Given the extra

---

[491]  Grafman and others contend similarly that Structured Event Complexes do not only support goal directed behaviors but also that "[s]tory grammar knowledge is an example of an SEC." Consequently, they propose to measure 'goodness of story narratives' in brain injury patients (Le, Coelho et al. 2011 119).

demand of global consistency, such benefits are more then welcome.

Now we did argue in the previous Part that there is continuity between conceptual and non-conceptual representations, as the former integrate as (component) representations those non-conceptual, modal representations. Simulation processes employ these not just for the reenactment of sensori-motor states but also for higher cognitive processes that employ language (Barsalou 2003 ; Barsalou 1999c ; Niedenthal, Barsalou et al. 2005).[492] This holds not only for single memory features or their inter-relations but also for the more complex Structured Event Complexes. Research with brain injury patients demonstrate that the impairment of processes that employ these SEC's includes their narrative capabilities, since the generation of a complete and well-structured narrative episode was found to be hampered in correspondence to their lacking action capabilities (Le, Coelho et al. 2011 119). Given the analysis of narrative simulation's functions for an agent's action performances, there is reason to expect that the conceptual format should facilitate his development of more complex and comprehensive structures, indeed.

With regard to narrative simulation, we found Ricoeur describing how through narrative emplotment "goals, causes, and chance are brought together within the temporal unity of a whole and complete action" (Ricoeur 1984 ix). More generally, we found that hierarchical – and heterarchical – action representations are prevalent in complex dynamic systems and their actions, for example in simple grooming behavior in flies (Dawkins and Dawkins 1976), or in more complex nettle leave eating actions of great apes (Byrne and Russon 1998), and in the implicit configuration of automatic actions in humans (Cooper 2003 ; Norman and Shallice 1986) However, the narrative emplotment that Ricoeur refers to is much more complex as it refers to hierarchical structures involving not just such single actions but practices, life plans and even the overall unity of an agent's life (Ricoeur 1992). Indeed are there developmental and patient studies that suggest a correlation between narrative capability and the capability for action organization and coordination. Before analyzing what resources are provided with the conceptual representation of action, let us take a short look at some of this evidence.

---

[492]  After a careful analysis of multiple lines of evidence in connection with three theses regarding types of groundedness of action cognition in motor abilities, the authors conclude that action cognition relies for both its acquisition and its constitution to some extent on motor abilities, but not completely. Consequently, some action cognition capabilities are not constrained by these abilities (Weber and Vosgerau 2012).

[493]  This is not the place to discuss preliminary developments that occur in child language learning and their effects on its interactions with its environment. Early development, for example, suggests that when a child is presented with verbal labels of multiple objects, this facilitates their categorization, inductive inferencing and their individuation capabilities, which is observable in its behavioral responses (Xu 2002).

Developmental studies do indeed demonstrate children's increasing capability for hierarchical organisation of representations in different domains like those of manipulation and speech.[493] An obvious explanation for this is that both depend upon shared cognitive and neural processes (Greenfield 1991).[494] With respect to their capabilities of narrative simulation, children develop structures in which increasingly more and more complex 'Goal-Action-Outcome' units figure. This development has been interpreted as being facilitated by the effective chunking or compressing of information with the use of these GAO units (Trabasso and Stein 1994) – which again confirms our previous arguments for the importance of chunking for effective and adaptive processing.[495] Apart from cognitive maturation it is the use of conceptual elements in narrative that contributes to these developments.

Since actions are not just composed of such relations between goals, actions and outcomes, but many psychological elements are involved as well, it is surprising how well children learn to understand, predict and explain human actions at an early age. An explanation for the acquisition of this skill is presented with the 'Narrative Practice Hypothesis' (NHP), developed by Hutto. Critical of the claims of theorizing accounts of action understanding and reviewing many developmental studies, the NPH purports that a child's sustained experience with narratives familiarizes it with the numerous factors involved in an agent's choice for a project: "[o]ften their reason for taking a particular course of action is influenced by their character, larger projects, past choices, existing commitments, ruling passions or unique circumstances and history" (Hutto 2007 35). Studies of mother-child interactions reflect this hypothesis and show that the complexity and contents of their joint story telling does influence the child's capability of understanding human action in psychological terms several months later (Turnbull and Carpendale 2009). What narratives provides the child with, so it seems, are coherent representations of both observable, indirectly observable – psychological, cultural – or even unobservable components that together allow it to make sense of actions. With these same representations, so we may assume on the basis of previous arguments, can it develop its own coherent narratives and multiple action plans. Let us not overlook the fact that this capability will not remain in place forever, unfortunately.

---

[494]  Meanwhile, evidence of involvement of Broca's area in facilitating hierarchical representations in action, language, and other domains supports this argument (Arbib and Bonaiuto 2007 ; Fadiga, Craighero et al. 2009 ; Hagoort and Levelt 2009 ; Koechlin and Jubault 2006).

[495]  There are several strategies that can be implemented in narratives for obtaining 'semantic reduction'. Comparison of narratives shows that simply applying Temporal connectivity to a story is less succesfull in such semantic reduction than the strategy of Action structure, with Causal connectivity occupying a middle position (Giora and Shen 1994). Hierarchical representation and chunking of information is indeed considered to be a most important form of 'problem solving' that narrative bestows upon human agents (Herman 2009).

Indeed, it has been observed in several studies that, conversely, aging is correlated with a decrease in the quality of narrative simulation. As the complexity of narrative simulation requires optimal executive functions and their declines in the elderly, their narrative tends to lose coherence, to be worse in integrating novel information and tracking multiple characters, to contain more irrelevant information, for example. From a study of adults of different ages, investigators conclude that narrative quality can indeed be taken as a general indicator of capabilities for cognitive and behavioral organization (Cannizzaro and Coelho 2012). Similarly, traumatic brain injury impedes narrative organization as was shown in a story re-telling test in which patients had more difficulty in providing their story with adequate structure. Again, impaired executive functions are held responsible for this, which was found in this study to be also correlated with decreased results in a card sorting task (Mozeiko, Le et al. 2011). After this sidestep to the fragility of the human capability for narrative simulation, we will take a closer look to the linguistic resources that are employed in forging coherent narratives while no longer subject to the constraints that are given with other representational formats.

When represented in linguistic format, an agent can easily represent actions irrespective of their temporal modality – whether past, intended future, or imagined actions – and freely 'act' upon these representations in many ways.[496] For example, given the recursivity that language brings along, an agent can indeed embed his action representation in other – linguistic – representations of actions or events, yielding him many new options.[497] When representing his action in an uncommon environmental context, for example, the agent is invited to consider whether new objects or tools could be integrated in the action, or to compare different actions via such representations.[498] Or he can ascribe a particular action intention to another

---

[496] In their target article on foresight or mental time travel, the authors even compare this capability with a theater production involving language-dependent contributions of a playwright, actors and a broadcaster. Acknowledging that mental time travel does not per se require language, they do emphasize that both MTT and language involve the "capacity to transcend the present in an open-ended and flexible manner" (Suddendorf and Corballis 2007 310). As a result, co-evolution of the two might have occurred. Investigating their development, Nelson confirms that MTT and language are interdependent, with an important role for cultural narratives in such development (Nelson 2007).

[497] Though Pacherie (Pacherie 2008) argues that distal intentions have to be made in a conceptual format, she does not explicitly take further linguistic features like syntax and recursivity into account, which are contributing to the benefits of representational redescription we mention here. Bruner explicitly mentions both the 'hermeneutic composability' and the 'narrative accrual' as characteristics of narrative that refer to the fact that narratives not only internally consist of nested components, but are generally also embedded in socio-cultural webs of narratives that contribute to 'narrative realities' (Bruner 1991). Such narrative recursivity knows hardly limitations once narratives are written and contained in 'external symbolic storage' and must no longer be contained as engrams in biological memory (Donald 1991).

[498] As Gerrig argues, 'the rich get richer' since those readers that were in the possession of relevant representations are better capable of integrating, maintaining and later recalling novel information as well (Gerrig 1993).

agent, challenging him to speculate about possible motivations or reasons that this person could have for it – potentially even sharing the intention with several others. Another option that presents itself when actions are represented as stories, is that they can become involved in what Shore refers to as 'analogical schematization' processes in which information and insights are carried over to different domains (Shore 1996), for example when a particular action is metaphorically applied with something else, like with performing an opera. In sum, in comparison to another medium for simulation like visual imagination, linguistically representing actions and intentions offers a wealth of options for redescribing these representations.[499] As a result, employing linguistic representation for narrative simulation of action, the agent can simultaneously embed it in a more comprehensive web of distal intentions while still enhancing the global consistency between these.[500] Indeed, there are highly specific linguistic features that are conducive to such consistency.

When developing a comprehensive hierarchical structure containing his web of intentions or action plans, the agent is in need of features that enable him to relate and connect action components, especially since it may be that not all components allow immediate integration in this structure.[501] All languages contain many words that can indicate one or another form of coherence between actions. Most effective are those words that allows an agent or observer to point out – or to question, for that matter – the causal connection between intentions and actions, like 'because', 'for', 'nonetheless', and so on. Evidence from several experiments demonstrate the cognitive benefits that using such connective words yield as they facilitate in subjects the understanding of a

---

[499]  Mental imagery can also have an impact on several behavioral or experiential measures in subjects (Jeannerod and Frak 1999 ; Kosslyn 2008). Yet mental images have certain limitations, different from linguistic or conceptual simulations. For example, subjects are unable to imagine ambiguous images, suggesting that self-produced images do not require further interpretation – which is different from percepts or from linguistic stimuli (Chambers and Reisberg 1985). Others have argued that reinterpreting – ambiguous - behavior is a typical human capability, dependent upon language (Povinelli and Barth 2005).
[500] Conversely, long term memory deficits have been associated with correlations between impairments in simulation tasks and speech tasks, which make patients produce results with less inter-item relations and less coherence in the events they've constructed (Romero and Moscovitch 2012). Generally, aging adults produce also less coherent and less efficient narratives than young adults. However, this effect is mitigated in aging adults with a larger vocabulary, which is probably helpful in constructing coherent and efficient narratives (Juncos-Rabadán, Pereiro et al. 2005). Even though much of the research reported on simulation either uses verbal stimuli or verbal reports, correlations between reported speech and simulation capabilities appear to be not just mere consequences of such experimental designs but to point to real interdependencies between these capabilities.
[501]  One could argue that the features mentioned here, or the indicators mentioned below, in part belong to metarepresentational content as they can contain information about the informational content of the action representation itself – for example by indicating the agent's belief or adherence to a particular representation component. It is argued that such metarepresentational content even need not be conceptual while still assisting an agent to improve and learn a representation in a more targeted way (Cleeremans 2006). Indeed, this can be carried further by arguing that the dopamine Reward Error Prediction signals are in fact carrying non-conceptual metarepresentational content (Shea 2012).

described series of events and induces better memorization.[502] Differences in response times demonstrate how such words help to raise expectations for future information and to strengthen relations between utterances (Mak and Sanders 2012). Other coherence indicators do not have clear parallels in observable action properties but are more unique to language, like words that indicate polarity– 'but', 'whereas' – or that present an additive relation – 'and', 'while'. The use of such words also result in facilitation of processing narrative simulation of actions (Knott and Sanders 1998).[503]

Since distal intentions are particularly complex due to their being extended into the future, the role of temporal structures in narrative simulation is particularly relevant. One of the main features of narrative is its use of linguistic resources to foster expectations and anticipations. Not just descriptions of intentions and actions are capable of instilling these, but even simple and canonized formulas about temporal structures ('In the beginning…') can do so (Gerrig 1993).[504] Indeed do behavioral results show that there are specific words which are facilitating the comprehension of such temporal structures in action descriptions. Naturally, such descriptions do not need to follow the rigid sequential order that actions must pass off in reality, which adds to their compactness (Zwaan 2008).[505]

Still other linguistic resources can help an agent's narrative simulation of his web of distal intentions to become more coherent while still reducing the amount of information that needs to be processed. Whereas observation or performance of actions might not reveal all potential differences between these regarding the attitude of the agent to these, their narrative simulation can do so. When an agent considers his actions and intentions, he can use many different 'indicators of self' to express his position regarding them for himself and others. There are several 'agency indicators' that allow him to can express in a differentiated way whether an action was voluntarily or not. Similarly, he can indicate his commitment to the action, its social reference and its evaluation of it. Moreover, he can also explicitly indicate the coherence of an action with his other commitments, evaluations and the like (Bruner and Kalmar 1998). Such

---

[502] This aligns with evidence that better hierarchical encoding of an action during observation by an observer correlates with his better recall and imitation of it afterwards (Zacks, Speer et al. 2007).

[503] When students were asked to develop deceptive autobiographical narratives, these narratives were found to be less complex and to contain less cohesion indicators. The investigators interpret their result as demonstrating that 'narrative distance' has an impact on the quality of narrative (Bedwell, Gallagher et al. 2011).

[504] Ricoeur emphasizes the fundamental role of narratives' temporal structures, as these often are not representing actions and events in chronological order, nor completely a-chronologically. Instead, different schemas for such structures are available (Ricoeur 1980).

[505] Zwaan argues in favor of a simulation theory along the lines of Barsalou's theory and asks for future research on "how mental simulations are orchestrated from moment to moment by the remarkable and often underestimated subtlety of human language" (Zwaan 2009 1149).

indicators can help both himself and others to judge this coherence and perhaps to modify it.[506] In that case those actions that are dear to the agent are put more central in the web of his intentions whereas others are put more at a distance. In that case, other indicators might need to be adjusted as well, inviting him to further sculpt the space of his actions.

As argued above, when an agent engages in narrative simulation he has at his disposal the representational format and resources that critically contribute to the representational redescription process that is implied in the explicitation of action representations. Consequently his actions, intentions and plans can be integrated in much more comprehensive representations without necessarily losing global consistency. Obviously, irrespective of the chunking and coordination involved in this process, narrative is still cognitively demanding, as testified by the increasing activation of particularly the right hemisphere in humans when completing a narrative – taken to indicate the efforts of synthesizing these narrative componentes into a coherent whole (Xu, Kemeny et al. 2005). It is therefore not surprising that there is a socio-cultural dimension involved in narrative simulation, which again facilitates and contributes to this process. Obviously, this dimension was already at stake in our previous discussion but we will focus more particularly on the socio-cultural nature of narrative in the next section.

### 4.2.5  The socio-cultural nature of some schemas for narrative simulation

It has been argued that the phylogenetic development of the modern human mind takes place in what can be characterized as a mimetic phase, in which public and communicative mimetic skills become increasingly important (Donald 1991).[507] Imitation and imitation learning of contents that are specific to a particular group or culture are playing an ever more prominent role in human lives since that paleonotological phase and are being associated with increasingly complex hierarchical representations of actions.[508] Such connection between mimesis, culturally specific

---

[506] Research in which subjects had to construct deceptive (pseudo-)autobiographical narratives shows that producing such false narratives is cognitively taxing and results in less linguistic complexity, less referential coherence and a greater distance between narrator and narrated self in comparison to truthful narratives (Bedwell, Gallagher et al. 2011).

[507] Donald's phases have been correlated with the ontogenetic development of children, in which narrative and the co-construction of narrative by children and parents does indeed play an important role (Nelson 1999). Later research confirmed that when children later construct their individual autobiographical narratives, they build upon the schemas that are socio-culturally available, which partly explains intercultural variability of the onset, quantity and quality of such narratives (Nelson 2003).

contents and intentional action has been at the centre of Ricoeur's work as well. In section III.4.1.4 we paused for a moment with his analysis of hierarchical levels of action, ranging from practices through life plans to the comprehensive unity of life, which were found to be to some extent determined by socio-cultural influences. The criteria of expert practices, or the nature of a parent's life plans, or the structure of one's autobiography are not made up by isolated and a-historical individuals (Ricoeur 1992). The same holds, Ricoeur argued in his volumes on 'Time and Narrative', for the narratives that humans tell, which generally comply to some extent with socio-cultural schemas with a long history (Ricoeur 1984-88). The use of such shared schemas again brings several benefits, both for an agent simulating his actions internally and for agent who are jointly simulating or discussing an action. The education of children consists partly of familiarizing them with such narrative schemas.

Indeed, according to the Narrative Practice Hypothesis that was mentioned in the previous section, children are being raised while engaging with their caregivers and others in narrative practice, which provides them with a shared basis for action understanding and narritive simulation of actions (Hutto 2007). More generally, narrative practice enhances not only the organization and consistency of actions of an individual agent as the exchange of such representations can also be considered as a collective cognitive activity that enhances the coordination between agents and their joint actions (Hutchins and Johnson 2009). Even in a simple perceptual task that two agents have to carry out it can be observed that dyads benefit from developing shared linguistic tools for their coordination, adapting to each other's way of talking, for example (Fusaroli, Bahrami et al. 2012). Reviewing literature from several lines of research, Tylen a.o. conclude that language can be a 'tool for interacting minds' bringing along four important benefits as linguistic representation: "extends the possibility-space for interaction, facilitates the profiling and navigation of joint attentional scenes, enables the sharing of situation models and action plans, and mediates the cultural shaping of interacting minds" (Tylen, Weed et al. 2010 3). Again, we can observe how benefits in two directions emerge: at the one hand, the space of options – this time for joint action - is enlarged with conceptual representation of action, while at the other hand language provides resources that help interacting agents to jointly constrain and determine a relevant sub-space of options, enhancing their consistent interactions.[509]

---

[508] A review and comparison of different lines of evidence, including the analysis of paleolithic stone tools, imaging studies of stone tool making and language processing, suggests that what connects the developments in stone tool making and language is their increasingly complex and hierarchical structures. The authors contend that these associated developments together will have affected the complexity of human intentional action and intersubjective learning generally, as can be derived from other archeological findings as well (Stout and Chaminade 2009).

As we've referred previously in this Part to research demonstrating the relevance of schemas or scripts, let us pause here for considering the relevance of culturally specific narrative structures.

Sculpting his space of actions such that it enables an agent to easily interact with other agents and his environment should be easier if it involves constraints that are not merely idiosyncratic but shared with his environment. This aspect of narrative is referred to in three of Bruner's listed ten narrative's features. The list includes the relevance of genres of narrative, with their specific structures and components, facilitating recognition and understanding of a recounted narrative. Furthermore, narratives contain normative elements having to do with cultural legitimacy of its contents and structures, which also constrain the expectations of both story tellers and listeners. A third aspect that merits mentioning and that is related to the other two is narrative's canonicity. Even though a narrated action must not always concur with a canonical action and can even breach it, the presence of shared canonical narrative structures does facilitate simulation and understanding of actions between subjects (Bruner 1991)[510] This facilitation consists mainly of the coherence between on the one hand individual intentions and actions and those of other agents and cultural institutions on the other, for which explicitly sharing, narrating, relevant representations seems to be an effective way.[511] Consequently, once could consider the cognitive act of story telling therefore primordially a social act and the result of this a kind of 'narrative entrainment' of different agents (Caracciolo 2012). In that sense, a narrative can be considered as a more elaborate and complex form of the template with open slots that were found to support sensori-motor skills in the beginning of this Part and that contribute to the process of chunking complex information and scupting the agent's space of actions.

Indeed, narrative – enabled by the features mentioned in the previous section – is

---

[509] Clark sheds light on the process of 'cognitive niche construction' for which humans use language and other means like spatial arrangements and tools. Important benefits of this niche construction are not just the expansion of options for action and interaction but also the reduction of complexity according to his account (Clark 2008).

[510] It appears that one can not only determine a limited set of narrative contents and structures that are prevalent within a particular culture, a brave attempt has even been made to classify a surveyable collection of narrative components which are largely shared between different cultures (Propp 2003). Ricoeur critically discusses Propp's 'logicization' and 'dechronologization' of narrative in (Ricoeur 1985).

[511] Trabasso describes how mothers' narration of stories to children functions as a scaffolding for their development of more complex action representations. Such narration employs socio-cultural models which help children to learn action planning while simultaneously socializing them (Trabasso and Stein 1994).

[512] Reviewing recent contributions to neurohermeneutics approaches to culture, emphasis was laid upon the different skills – ranging from motor to symbolic practices – associated with a particular culture (Winkelman 2003).

said to "provide[...] templates for behavior in physical as well as moral-cultural worlds" (Herman 2003 182).[512] Not surprising is the fact that the notions of script and schema, discussed above, have been introduced with reference to socio-culturally specific actions like coffee making (Cooper, Schwartz et al. 2005) or visiting a restaurant (Schank and Abelson 1977). Such schemas contain complex action representations in which several components actions, agents, and environmental props figure, which are interconnected through expectations, dependencies and the like. On a more broader scale and in the wake of the 'cognitive turn' that the social sciences have made in the last decades, sociologists and anthropologists have taken an interest in the cognitive effects of socio-culturally shared representations (Shore 1996). Although we've observed in our discussion in Part I of Marr's three levels of analysis that the algorithmic implementation of a particular task and its neural implementation cannot be derived unambiguously from each other, changing a task's representational format often does influence the cognitive and neural processes required for its performance.[513]

Do cultures indeed differ with regard to those higher levels referred to by Ricoeur as practices, life plans and the overall unity of an agent's life (Ricoeur 1992)? In a review of 'neuroanthropological' research, reference has been made to the fact that the Structured Event Complexes, to which we've referred repeatedly in this Part, are in many cases determined by socio-culturally specific contexts and actions (Dominguez Duque, Lewis et al. 2009). With regard to practices and life plans, the differences in contexts and norms between cultures are more prominent and have been implicitly present with regard to the levels of motor intentions and proximal intentions.[514] However, with regard to the unity of an agent's life, it seems to be more difficult to investigate how socio-cultural influences can modulate the cognitive and neural processes associated with his self-representation. Perhaps still with serious limitations,

---

[513]  Reviewing cultural neuroscientific results, the authors conclude that cultural differences can have not just a functional but even a lasting structural impact on brains (Han and Northoff 2008). It should be noted, though, that comparing cognitive and neural processes in subjects from different cultures brings along some fundamental conceptual and scientific challenges (Roepstorff 2013 ; Roepstorff and Frith 2012). Nonetheless, there is increasing evidence that cultural differences do influence both cognitive, behavioral and neural processes in subjects (Ames and Fiske 2010 ; Chiao, Cheon et al. 2013 ; Choudhury and Gold 2011 ; Dominguez Duque, Lewis et al. 2009 ; Losin, Dapretto et al. 2010 ; Nisbett and Miyamoto 2005 ; Vogeley and Roepstorff 2009).

[514]  Taboos, for example, are present in all cultures and predominantly regulate bodily interactions and food. Although it may be that many food taboos can be related to experiences of disgust in connection with particular – rotten – foods, it is obvious that additional social and moral values and norms have become associated with such foods. As a result, the representation of a taboo can become 'enriched' with normative, social and religious components that subsequently engage many different cognitive – and social – mechanisms (Fessler and Navarrete 2003). As a result of such processes, taboos can become strongly generatively entrenched and have a wide-ranging impact on an agent's web of intentions and action plans.

there has been developed a line of research that can offer us some preliminary insights in this.

There have been made several investigations into differences in self-related processing or self-referencing between subjects from different cultures. Although it must be granted that the tasks involved in such investigations are relatively simple and do not amount to narrating an autobiography, neural, cognitive and neural evidence does suggest that the representation of self in relation to others is not uniform across cultures. For example, correlating with the broad distinction between a more collectivist Asian culture and a more individualist Western one, students from different backgrounds engaging in self-referencing categorization tasks were found to display different cognitive and neural response. Cognitive and neural responses to mothers and unknown persons were clearly distinct from responses to self in Western students, whereas in Chinese subjects the line was drawn between mother and self versus unknown persons (Zhu, Zhang et al. 2007).[515] Intriguing results with Chinese buddhists compared with Chinese christians suggest that the long-time exposure to religious narratives and practices differing in terms of their self-focus or no-self doctrine, respectively, does influence cognitive and neural responses to the self (Han, Gu et al. 2010).[516] However, other research demonstrates that such cultural differences in self-representation are modifiable, for example by individual endorsement of cultural norms and – implicitly – by priming (Chiao, Harada et al. 2010). More directly, self-identification with a particular race results in larger empathic responses with other racial group members and increased activations of a large set of neural areas with that is involved with more than only empathy (Mathur, Harada et al. 2012).[517] In fact, as we will see in the next section, self-representation or self-identification is correlated with a large neural network which appears to play a more generic functional role instead of having a specific function.

---

[515] Although humans may share a 'trans-species core-self' responsible for fundamental self-referential processes with other animals, this does not stand in the way of its being sensitive to environmental and social influences (Panksepp and Northoff 2009). Indeed, this concurs with our argument that dynamical and adaptive systems are capable of integrating environmental information in some way or another.

[516] The prevalence of a hierarchical or more egalitarian societal norm also influences the perception of others and the distinction between out- and in-group members. This has an influence on cognitive and neural empathic responses, with subjects from a more egalitarian society displaying less distinction between out- and in-group members (Cheon, Im et al. 2011). In this context, too, manipulation of the cognitive strategy of subjects – for example by demanding different types of categorization – can modulate such responses on all levels (Sheng and Han 2012).

[517] There are several cognitive strategies available which influence the perception of and empathic responses to differences in group membership. Mirror neuron systems are involved in these responses and it appears that their activations can be modulated via the use of such cognitive strategies, contradicting simplified statements of the 'hard-wired' or 'innate' nature of our social, empathic brain, as we've argued elsewhere (Keestra 2012).

To resume the previous sections, socio-cultural norms and representations can have an impact on behavioral, cognitive and neural responses, confirming the relevance of narrative - containing such socio-cultural contents - for an agent. The agent's structural self-perception can even be affected by it as other research shows. Most agents tend to think of themselves as  being autonomous, integrated and separate from other selves. However, as is evident from many foundational narratives, in some cultures a more porous and disembodied characterization of self is prevalent, making dissociative phenomena like trance or possesion acceptable and not dismissed as pathological. Such cultural narratives appear to influence automatic and controlled cognitive and neural processes like those underlying attention, perception and emotion, related to such dissociative phenomena (Seligman and Kirmayer 2008).[518] The narrative that an agent, who is part of such a culture, will develop about himself will likely display many differences compared to another narrative in which his interactions does not include interactions with spirits of ancestors and the like.

Given the mutual influences between socio-cultural narratives and the individual agent's cognitive processes, it has been argued that we should not be surprised that the representations involved in such narratives tend to be reproduced in a relatively stable way, even though they undergo modifications throughout cultural history (Sperber and Hirschfeld 2004).[519] Such characteristics of the socio-cultural transmission of representations may rely particularly upon the hierarchical structure of these representations, again. Several lines of evidence suggest that humans are subject to a 'hierarchical bias' as during transmission processes the higher, more abstract levels of action representations tend to become increasingly important whereas the proportion of lower level information decreases (Mesoudi and Whiten 2004). Given the potential for abstract representations provided by language, this may explain why narrative structures can spread relatively easy and maintain stability over such long periods of time, influencing the behavioral, cognitive and neural processes of large groups of agents.

Having started these sections with the relatively simple simulation of a single future action, we have now arrived at underscoring the importance of the narrative simulation of the agent's more complex web of intentions. At the end of this last Part, let us take

---

[518]  Attention could also be paid to the use of psychoactive drugs in some shamanist cultures, further enhancing the experiences of trance and possesion and influencing neural processes (Whitley 1998).

[519]  Sperber defends a view of a 'massively modular mind' and contends that his notion of an epidemiology of representations depends upon this view of the mind (Sperber 2005). Apart from the fact that he seems to take the notion of modularity even stricter than Fodor required (Fodor 1983), our account of the brain as a complex and dynamic mechanism that is capable of kludge formation can explain such epidemiology without implying such – problematic - modularity.

a final look at some empirical studies concerning the implementation of simulation – particularly narrative simulation. An important question is whether we will find kludge formation at this stage, or not. Given the pertinence of narrative to the agent's wider web of intentions and action plans, it seems difficult to expect the formation of a particular (component) mechanism responsible for processing such complex and multifaceted representations. Moreover, Ricoeur has emphasized that the narrative of an agent's life is never finished but requires continuous recounting and modification in view of his new or revisited experiences and actions (Ricoeur 1991b ; Ricoeur 1992). Implementing this in a particular mechanism appears to be quite impossible, or is it?

### 4.2.6  Narrative simulation and some evidence for its implementation

After considering motor, proximal and simple distal intentions, we have now even touched upon narrative simulation as another process in which action representations are involved. We have defended how over time an agent's space of actions is sculpted as his growing expertise with particular actions leads to the formation of kludges, correlated with relevant representations of his expert actions. Whether relatively simple templates of sensori-motor representations or the more complex Structured Event Complexes, for example, his expertise results in flexible yet fast responses in accordance with his established intentions, even without exerting conscious control. Our plea for extending this framework to distal intentions may have caused some wonder, as distal intentions are usually considered to be articulated consciously and rationally. However, we can still ask: are such distal intentions perhaps also capable of becoming generatively entrenched in the complex mechanisms that are underlying an agent's cognitive and behavioral responses? Can distal intentions contribute to a sculpted space of actions in such a way that conscious and rational decision making is not always required for these intentions to influence his actions? Of course, we are not defending a position which holds that an agent's narrative simulation of his actions will always and comprehensively control all his actions.[520] Nonetheless, as we've argued that sculpting the space of actions is a process subject to multiple influences, it is still relevant to consider the implementation of narrative simulation. Given the complexity of the task of distal intention formation and narrative simulation, we may expect their implementation to be rather complex, too.

---

[520] Indeed, for improving an agent's moral behavior it is usually not sufficient to improve his narrative, perhaps in part because of a lack of overlap of the neural processes that underlie our narrative capacities and those involved in action planning and performance (Bickle 2003). On the other hand, there are several cognitive strategies that involve narrative simulation and have an impact on action, like implementation intentions, counterfactual reasoning.

Before looking more closely at such implementation, we will consider some indices that cognitive decline or other pathologies are often associated with impairments of narrative and of action. This concurs with our earlier observation, for example in section III.4.2.4, that the domains of action and speech share some cognitive and neural processes. For example, patients with Alzheimer's dementia are increasingly losing the capability of generating the lower level details of action representations in both domains (Addis, Sacchetti et al. 2009). Studies suggest that it is the decrease in goal-directed executive function in ageing adults that correlates with changes in their narrative's structure, which becomes less coherent, less informative, less complete and more confusing, probably due to less succesful implementation of organizational, hierarchical schemas (Cannizzaro and Coelho 2012). Similarly, schizophrenic patients, characterized among others by their action disorganization, are making less action plans and have more difficulties in simulating the details of future events (de Oliveira, Cuervo-Lombard et al. 2009).[521] Such declines appear to be foremost a result of impaired capability of developing comprehensive and complex action representations. On the other hand, these same patient groups suffer from impaired self-experience and self-concept, concurring with the fact that in narrative simulation is essentially about an agent and his actions.

Concurring with this patient evidence are studies with other patients, for whom narrative contributes to improvements in mental and physical health. Patients suffering from traumatic experiences were shown to benefit from writing exercises. Especially those patients who created increasingly coherent narratives about their experiences – indicated by their use of insight and causal words – improved significantly (Pennebaker 1993).[522] Extending this finding with a review and experiments, Klein argues that it is particularly the increased coherence of patients' narratives that is responsible for such improvements. Such change in coherence usually entails the transformation of the mental representations that patients have, also modifying the stress-related components. Furthermore, narrative coherence also limits the ability of intrusive memories to disturb patients (Klein 2003).[523] Although Ricoeur may not have had these patients in mind, the evidence concurs with his emphasis of the importance of

---

[521] As we will elucidate shortly, several processes are involved in narrative simulation. In schizophrenic patients, it is suggested that their deficits in reality monitoring and in strategic memory retrieval contribute to their difficulties in simulation tasks (Raffard, D'Argembeau et al. 2010).

[522] Such evidence confirms the 'Immersed Experiencer Framework' which emphasizes parallels between real-world experiences and text-processing, since in the latter situation a simulation – along the lines of Barsalou's theory (Barsalou 1999c) - of the former is established (Zwaan 2004).

[523] In her review of available evidence of the beneficial results of expressive writing for patients, Klein mentions also unexpected results like reduced blood pressure, improved immune function, and improved working memory (Klein 2003).

narrative for the agent's life: "If my life cannot be grasped as a singular totality, I could never hope it to be successful, complete" (Ricoeur 1992 160). Let us first consider the implementation of narrative, before focusing more in particular on the potential implementation of self-representation, which plays such an important role in narrative.

According to Mar's review of the literature (Mar 2004), the comprehension and production of narrative relies on many different neural areas, recruited for three broad component cognitive processes: memory encoding and retrieval, the integration of information in order to create coherence, and further elaboration or simulation. Consequently, the responsible neural networks must include at least hippocampal and working memory areas, for the first process. Depending on the kind of information that must be integrated, a wide range of processes are candidates. In any case, a representational structure like Grafman's Structured Event Complex must be employed for maintaing information and integrating further information in it. Simulation of predictable events requires processes that enables an agent to draw inferences or raise expectations. These last two processes rely on large prefrontal areas, responsible for ordering and selection of contents (MPFC)[524] and for constituting temporal order and offering working memory (DLPFC). Moreover, temporoparietal and temporal regions are involved, contributing to mentalizing processes, in which MPFC is also involved. Finally, the posterior cingulate cortex appears to be recruited for autonoetic awareness, enabling the narrator to truly experience the narrated simulation, including its affective aspects (Mar 2004).

However, not all evidence converges with this picture and more detailed investigation and interpretation of specific cognitive contributions to narrative simulation still stand out, as is the case with their neural implementation. As an example, we will mention how several investigations help to specify the processes recruited for coherence building in narrative, even though they do not all point into the same direction. Imaging subjects during reading comprehension tasks did suggest that DMPFC is involved in understanding a narrative's coherence. Yet other studies reported DMPFC activations also in incoherent sentence conditions as well as in theory of mind tasks, suggesting that those activations are contributing to still unidentified cognitive tasks in narrative comprehension (Yarkoni, Speer et al. 2008). More generally, imaging experiments with subjects performing comprehension tasks at different levels of narrative – ranging from words via sentences to narrative – showed that the role of both hemispheres can be somewhat differentiated. It appears that particularly for forging a coherent

---

[524] This concurs with results in which MPFC activation appeared to be recruited for mediating attention in the simulation of distal intentions (Okuda, Gilbert et al. 2011).

representation of all narrative components, right hemispheric activation is required (Xu, Kemeny et al. 2005).[525] As could be expected from the contents of a narrative – consisting of the synthetic emplotment of heterogeneous elements, as we've learned above – its implementation is still contested area. What, then, should we expect with regard to the implementation of the agent's self-representation in narrative? Given its rich associations with memories, intentions and plans, emotions and motivations, expertise and knowledge, it seems implausible to expect an implementation that can lend itself to a form of kludge formation, as well. However, we will suggest that there are indications of the latter, indeed.

As mentioned earlier in the context of the simulation of simple distal intentions, the default-mode network (DMN) plays a role in such simulation (Buckner and Carroll 2007 ; Schacter, Addis et al. 2008). For some time, it has been assumed that this DMN is actually anti-correlated with task-related activities in the brain, leaving the functional role of it undecided. But this dichotomy has become a matter of debate (Fox, Snyder et al. 2005). Indeed, even in simple cognitive tasks that are 'perceptually decoupled' from the immediate context, this DMN is recruited. However, this DMN recruitment does then affect negatively the goal-directed actions that the agent is performing at the same time – a negative effect that does suggest a distinctive role for DMN for tasks that do not focus on present external, environmental goals (Smallwood, Brown et al. 2012 ; Smallwood, Tipper et al. 2013). So it is particularly with regard to tasks that require some internal focus and self-representation, like in distal intention and narrative simulation, that we should expect DMN activations, as it is held to be involved in maintaining information and its processing in interpretive and predictive tasks (Raichle and Snyder 2007).[526] It should be noted in this context, again, that these tasks do not only involve the representation of relevant information, but also some meta-representational capability involved in its evaluation and interpretation. DMN involvement in tasks requiring these capabilities is confirmed by different lines of research. We may consider this complex yet distinct network to be a kludge that is recruited in several mechanisms that are responsible for some important cognitive tasks.

For example, DMN is activated when agents were asked to simulate how they would solve goal-directed problems sometime in the future. In addition, however, executive regions were recruited, reflecting probably the representations that are relevant for the

---

[525]  In addition to the processes contributing to narrative simulation there may be a role for processes that check for the coherence or accuracy of narrative.

[526] Nonetheless, a review of DMN activations in patients suffering from a variety of disorders suggests that interferences of DMN activities with external goal-related tasks are correlated with several disorders (Broyd, Demanuele et al. 2009).

problems at stake (Gerlach, Spreng et al. 2011), like the Structured Event Complexes discussed earlier. Also in less specific simulation tasks, DMN is recruited, as when agents are challenged to engage in narrative fiction, in autobiographical narrative or in mentalizing tasks (Spreng, Mar et al. 2009). It might be speculated that it is indeed the synthesizing into a coherent narrative of heterogeneous types of self-related information that relies upon such DMN activations. Such speculation receives initial confirmation from an imaging experiment demonstrating particular involvement of DMN towards the end of a narrative, when the task of its integration is particularly relevant (Egidi and Caramazza 2013). Such recruitment during narrative seems to be implied because of the rich connectivity of this network with cognitive, affective and mentalizing systems. Concurring with this is the fact that when a subject has difficulty with particular narrative tasks, his DMN network is modulated accordingly (Wilson, Molnar-Szakacs et al. 2008).[527]

Even if the DMN is relatively de-activated when ongoing tasks are activating specific networks, such experiences should at times have a subsequent effect on it. Indeed, the agent's experiences may at times need to be integrated in this network as it is involved in processes that are relevant for his forming distal intentions and narrative simulation. There is recent evidence that such 'updating' of DMN is taking place after a series of learning sessions of a particular tasks, showing a correlation between DMN modulations and the results of learning (Taubert, Lohmann et al. 2011). Even more recent is evidence that after just a single training session, with neurofeedback assisted performance of both simple voluntary gaming tasks and tasks requiring subjects to simulate a future research project, DMN components show results of Hebbian learning, leading the authors to speculate that: "the resting-state patterns may constitute a powerful brain-wide and personalized 'window' into the personal history of brain activations in individual subjects" (Harmelech, Preminger et al. 2013 9496). Per implication, it would also explain the important role of the DMN for narrative simulation.

Needless to say, narrative simulation requires the integrated representation of many different types of information and may range from single distal intentions up to the narrative simulation of the unity of life – the highest level of action as mentioned by Ricoeur (Ricoeur 1992). In such simulations, the agent's self plays an important role. A subset of DMN components appears to be specifically engaged when an agent is representing, monitoring, evaluating and integrating information that is related

---

[527] Another interesting finding is that in chronic pain patients, there appears to be a correlation between their accompanying symptoms like depression and abnormalities in decision-making, and disturbed dynamics of their DMN (Baliki, Geha et al. 2008).

to himself, for example when this information is judged to hold of him as a person. These cortico-midline structures (CMS) are found to be consistently activated in such processing self-related information stripped of contextual information, and is interpreted as being partly responsible for the experience of a stable self, a 'core self' (Northoff and Bermpohl 2004).[528] Indeed, particularly when subjects were asked to engage in internally focused and not externally focused self-related simulations, these CMS showed subsequent modulations (Schneider, Bermpohl et al. 2008). We might expect from this that t is important to be able to deactivate this network in case of engagement with externally focused tasks, presented by one's environment. A study with patients with major depression does indeed show this balance to be disturbed. These patients exhibit an increased self-focus and decreased external focus, associated with feelings of hopelesness, apathy and ruminating. Correlated to these symptoms, abnormalities in activations of their (subcortical-)cortical midline structures have been found (Grimm, Ernst et al. 2009). As important as both the default-mode network and its network of cortico-midline structures are for typical simulation tasks and as much as these are involved in updating representations of self and important self-related information, an agent must be able to disconnect or leave out such kludges from the cognitive mechanisms that are invoked when he is required to interact fast and flexibly in his environment.

So given the requirements of updating and evaluating both self-related and contextual information, it is understandable that our expert singer needs a few hours to reflect upon his day of rehearsals. Indeed, there is a lot that he has experienced and learnt during a day in which he had to repeat several times the difficult fast run in Don Giovanni's invitation-aria with a conductor who has a tendency of speeding up the tempo, in which he had to get used to the new stage props and the distances he had to cross singing on the stage, and in which he was again confronted with the stage director's request of behaving like a rapist towards Zerlina. During those resting hours he does what most agents do after an intense day of activities: engage in some mind-wandering along several situations of the past hours, pausing at specific moments and then zooming in on particular elements of a situation or of an interaction or action, or trying to remember what the conductor or another singer has said.[529] Some particular content of his thought does attract his attention and invites him to reconsider his

---

[528] Somewhat extended, subcortical-cortico-midline structures are considered responsible for a core-self that is taken to be present not just in humans, nor perhaps only in mamals but in other species as well (Panksepp and Northoff 2009).

[529] It has been hypothesized that activation patterns during sleep reflect the replay and reconsolidation of memories related to performances during the day (Pennartz, Uylings et al. 2002). Similarly, during rest periods such reactivation occurs, correlated with learning effects in humans (Daselaar, Porat et al. 2010).

performance or response, which at times he immediately repeats – sometimes silently, sometimes sitting for a moment at the piano –with an eye on modifying or improving it. Fortunately, in most cases he is assured that he must only slightly adjust (the open slots of the templates underlying) the automatisms or schemas, which is generally easy. Other moments require him to engage at different types of processing, of course.

An issue that has been nagging at him and has now again raised doubts, is his decision with regard to the director's request: should he, or should he not act as a rapist, even though he has always tried to avoid agression in his behavior towards the other sex? How would his decision cohere with his other performances or the statements he has made in interviews or private conversations with regard to gender issues or to the role of opera? For he realizes that his hesitation stems from some immediate yet unarticulated disturbance that he felt when he initially rejected to do so. It would be odd, he now realizes, if he would dismiss the director's suggestion out of hand just because he rejects agression towards women. For he would of course be one of the first to realize the difference between theatre and reality. But it was only today, when they were also rehearsing the final scenes of the opera, that the director demanded that Don Giovanni was to die not just by the hand of the Commandatore's marble hand, but also by the helping hands of all women that the Don had harrassed during the opera – by the hands of Donna Anna, Donna Elvira, Donna Elvira's maid and by the violent hands of Zerlina. So when our expert singer now looks upon the opera as a whole and the fate of his role, he smiles upon the poetic justice eventually wreaked upon him. Satisfied with that outcome – and perhaps also a little bit enticed by the agression inherent in the role – he decides to accept the stage directions and immediately walks to the piano and starts exercising 'La ci darem la mano' with a somewhat different tone of voice and tempo. Having had already some phantasies of playing the scene with Zerlina with more agression, some phantasies even disturbing to him, he does not need to search long for an appropriate tone.

What our singer probably will never know, is that the stage director had only recently decided to have the women take part in Don Giovanni's death as he wanted to keep our expert singer with his beautiful voice and convincing playing aboard, knowing that he would not accept to enact a raping scene if it was not compensated by some retributive justice. Together, yet without explicitly discussing it, singer and director contributed to the overall harmony and coherence that should characterize a perfect opera – opera meaning 'works', 'labours' or somewhat loosely: actions.

# CONCLUSION AND SUMMARY. WHY SCULPTING THE SPACE OF ACTIONS MATTERS[*]

Why should we praise someone for performing so well, even though we usually reserve praise for consciously deliberated and chosen actions and less so for actions that appear to be produced automatically and effortless? Observing such action performance by an expert singer performing an opera role or a seasoned citizen engaging in moral action, one can easily fall prey to this paradox of expert action: instead of praising the expert, shouldn't we rather praise the novice, even though he may not be performing equally well and smoothly because he is at least deciding about his actions step by step? However, if we were to agree with this position, it would amount to admitting that during the process of acquiring expertise or skill, a person loses his – or her – admirability. For in proportion to his increasing expertise, his performance depends less and less on conscious and direct control of action. If indeed we take such immediate, conscious action control as a necessary ingredient of any form of intentional action, then we may be forced to withhold our expert singer the capability of intentionally executing his complex performance.

Such paradoxes have bothered philosophers, scientists and laymen alike since ancient times, trying to understand and explain human action. Indeed, Socrates aimed to avoid this paradox by positing that it is by definition through reasoning that an agent determines intentional and voluntary action. Aristotle clearly rejected this position, arguing against a simplified theory ('logos') that is at odds with our experiences and attitudes regarding this phenomenon and our reflection upon them (Ethica Nicomacheia 1145 b 27-28). In order to accommodate this, Aristotle added two elements to his account that allowed him to propose a more satisfying account. First, he recognized that human action is characterized by causal pluralism and not just determined by a single cause. Second, he realized that it may be necessary to carefully redefine current concepts or even to introduce additional concepts if paradoxes and inconsistencies arise within a theory of action.

It is such a navigation between conceptual and empirical insights that is undertaken in this dissertation, too. We wanted to do as much justice to the differences between expert and novice action as to their continuities. These differences are not only observable in the greater complexity, speed and flexibility of expert action in a given situation. In addition, an expert is generally better able to intentionally plan, organize,

---

[*] On pages 371, 373, and 375, figures I, II, and III offer simplified representations related to the arguments made in Parts I, II, and III respectively.

modify and describe his action than a novice. Notwithstanding these differences, an expert didn't change his brain or body so we must explain how development and learning have enabled the same body and brain to produce a strikingly different performance.

For this explanation we have introduced and elaborated on the concept of 'sculpting the space of actions' as an explanatory tool (see section III.1.1). This concept allowed us to develop a comprehensive integration of interdisciplinary insights in (the emergence of) expert action, particularly philosophical and cognitive neuroscientific ones. In order to understand and explain how an agent determines his action in a given situation, we propose to consider it as a problem of finding a suitable candidate from the large number of actions that he could potentially perform. We propose to represent all of his action options as separate locations or subdomains within a multidimensional 'space of actions', specific to the agent. Each action option is located somewhere in this space of actions, its specific location being defined by numerous factors. Some action options are represented more prominently than others, occupying a larger sub-space at a more central location in the space of actions and therefore having a bigger chance of being selected and performed.

The space of actions that each agent has is not static. Instead, we argued that it is 'sculpted' in several ways, with both long-term and short-term effects. Long-term and stable changes that happen to an agent's space of actions are results of his development and learning. Such changes obtain when new actions are added to it, when well-practiced actions gain in prominence within the space, when unlearned actions are relegated to small and peripheral locations. Due to these long-term changes the options are no longer uniformly distributed in the space of actions. Instead, the experienced agent's space of actions is constrained in many different ways instead. According to this concept, when an agent is acquiring a skill or is gaining in expertise, his space of actions is subject to a sculpting process that affects particularly the sub-space of actions belonging to the domain of expertise.

We argued that when explaining an agent's action performance in a given situation, we should acknowledge that this process of sculpting the space of actions occurs in the short-term as well. For even though an expert's space of actions is sculpted more than a novice's, the selection of a particular option for performance is still subject to the conditions of the particular situation he finds himself in. A mountaineer who fell in see will be less inclined to climbing than to swimming, an opera singer must do his best to act Don Giovanni-like charming to a detested colleague: external and internal conditions further sculpt the dimensions and structure of his sculpted space of actions in a more transitory sense during the action itself.

This concept of 'sculpting the space of actions' as an explanatory tool resulted from the preceding investigations made in this dissertation. First, in Part I, we critically discussed different methods of explanation used in cognitive neuroscience, looking for an explanatory method that can integrate both the causal pluralism and the effects of development and learning in an account of expert action. In Part II we then proceeded by applying the selected explanatory method to different theories of development and learning, accounting for both stable and dynamic effects of gaining expertise more generally. Finally, in Part III we turned to the explanation of human action and expert action in particular. Building upon our insights regarding explanation and about development and learning, we integrated philosophical and cognitive neuroscientific insights in it. This integration was facilitated by adding the concept of 'sculpting the space of actions' as a valuable explanatory tool. In the remainder of this section Conclusions and Summary, we will concisely travel after these navigations.

Part I was devoted to a discussion of four different methods of explanation pertaining to the field of cognitive neuroscience. All four methods offer solutions to how we should gather and integrate insights from neurobiological, computational, cognitive psychological and related studies in such a way that they together allow understanding and explaining complex phenomenon like intentional action. Complicating factor is that cognitive phenomena have proven difficult to define and without a clear definition it is unclear whether presented insights do in fact apply to the same phenomenon. Philosophical analyses can help with such conceptual matters but the four explanatory methods showed that they propose quite different relations between conceptual and empirical investigations.

The method discussed in chapter I.2 prescribes a crucial role to the way philosophers carry out conceptual analysis of a psychological function like consciousness or emotion. Its authors, Bennett and Hacker, maintain that empirical studies can only be usefully done on the basis of a clear definition that is reached through such an analysis. We pointed out how they assume that even for a notably complex function like consciousness, it is possible to develop a consistent conceptual framework, allegedly based upon an analysis of the concepts commonly applied to it and the behavioral criteria associated with it. This assumption was found to be unwarranted with regard to consciousness, or regarding a conceptual divergence like 'blind-sight' into account. Instead of rejecting such surprising concepts, as the authors propose, we defended that they be used as heuristics pointing us the way to unexpected interactions between functions or to a causal pluralism that has gone unnoticed. In other words, we argued

against strictly separating conceptual and empirical studies and for using insigths from one as a constraint or heuristics for the other's investigations.

More productive is the method proposed three decades ago by David Marr and influential in cognitive neuroscience ever since, treated in chapter I.3. We found that it concurs to some extent with the previous one in that it assigns an important role to what is called the 'computational theory' or task analysis pertaining to a function, like vision. This computational theory should provide us with insights in the function's goal, taking into account also the function's role for other functions or in a wider context. This method strongly diverges from Bennett and Hacker's in that it explicitly prescribes how scientists should develop two more theories to gain a more comprehensive insight into a function. The 'algorithmic theory' explains how the information used for a task is represented and transformed, with usually several options available. Although Marr maintained otherwise, we found him using all three theories to constrain each other. For example, a particular task can theoretically be carried out with different kinds of representations, yet based upon the brain's neural properties one kind is more probably used than another kind of representation. It is by such an integration of insights, applied to its various objects, that cognitive neuroscience can make progress, as we defended throughout this dissertation.

Chapter I.4 was devoted to the method of explaining consciousness by looking for its neural correlates. We pointed out that this method does neither require a conceptual analysis nor a task analysis as it accepts that there is no generally accepted definition of consciousness available. A similar permissiveness was found with regard to it's expectation that a particular conscious state should be 'mapped onto' a particular neural state, without prescribing the sort of relation between the two. However, notwithstanding their liberal stance, we found that researchers still cannot avoid differentiating between studies by using – sometimes implicit - concepts of consciousness. Alternatively we found how they created coherence between studies by looking for overlapping neural correlates, assuming that these findings do indeed pertain to the same object of study. Morever, a particular neural process was presented as a defining criterion of consciousness. However, it still remains to be determined how this neural process contributes critically to consciousness, which is impossible without at least a preliminary definition or task analysis of consciousness. Determining the contribution of a neural process to it would then require formulating a computational theory or an algorithmic theory in Marr's terms for it, so we argued. Thus, investigating the neural correlates of a particular function in a fairly liberal way may indeed be useful, but only as a first step.

Chapter I.5 finally argued that the method of 'mechanistic explanation' facilitates

the required integration of insights better than the methods discussed so far by dividing the task of explanation of a function over many different perspectives and offering the means for their integration. It requires the application of a few heuristics for this task division and enables scientists to reconsider and adjust the formulated 'explanatory mechanism' in light of subsequent results. These heuristics are: the definition of a cognitive function, its decomposition in component functions, and finally the localization of these component functions in the organism and its brain. Each of these steps can be iterated in light of newly gathered insights or applied to further subcomponents. Memory, for example, has been defined as not just the storage but also the retrieval of information, as studies show that these can be differentially influenced or lesioned. With such a redefinition, the decomposition of memory has also changed and consequently, additional localizations in the brain have been scrutinized. We noted that developing a mechanistic explanation for a (component) function also benefits from formulating the three theories prescribed by Marr: what is the task of this particular function, what representations and transformations are involved for it, and how is it neurally implemented?

In addition to the fact that mechanistic explanation enables the integration of different insights, it is the only method that provides the resources needed for explaining the effects of development and learning which is particularly relevant for our project. For this, we developed here four different kinds of modifications that an explanatory mechanism can undergo, affecting the number and configuration of its components and also the interactions with its environment. Even though we acknowledged some limitations of this method of mechanistic explanation, we concluded that this method was most useful for the task at hand: explaining human expert action as being produced by a complex interaction between mechanisms and intentions.

Part II shifted to discussing several cognitive (neuro-)scientific theories about development and learning. Its aim was to consider whether we could apply the method of mechanistic explanation to these theories. It started with a preliminary general observation that development and learning generally lead to structural and stable changes in a mechanism responsible for a particular function. Because of their stability, such changes can accrue as they build upon previous changes, contributing to the hierarchical structure that complex and dynamic mechanisms usually have. As a result, earlier changes tend to become ever more deeply 'generatively entrenched', in Wimsatt's words, in the mechanism that subserves a particular function: a change has stable effects on the responsible mechanism and these effects are subsequently

involved in its further developments. We referred to such changes as cases of kludge formation, affecting both the structure and workings of the mechanism. Seven general kludge characteristics were set out, some of which appeared to be useful in our subsequent discussions of the theories of development and learning. Important, for example, was that as mechanistic explanation aims to elucidate an observable function, kludge formation must initially be characterized in functional terms. From this functional characterization we can unfortunately not directly derive a specific algorithmic theory nor a specific theory about its neural implementations, as was noted earlier. Indeed, it may be possible that differences between individuals can be found with regard to the representations or neural processes involved, even though these differences do not show up in their performances. A final kludge characteristic referred to the integration of environmental information in a function's explanatory mechanism. This explains why cultural differences can have a stable impact on it and not just on observable performances.

The first theory of development and learning, discussed in chapter II.2, was neuroconstructivism. Although focusing primarily on Karmiloff-Smith's work, which distinguishes between the stages involved in the acquisition of skills and expertise in children, we also applied this theory to adult learning. We found that neuroconstructivism assigns an important role to the process of 'Representational Redescription' that is involved in learning, concurring with the importance of algorithmic theories in cognitive neuroscience. During learning, the representations involved in executing a task do not remain the same but gain in complexity and structure, becoming increasingly available to the learner for explicit adjustment and correction, as when a singer learns to fathom the structure of his music score. Next to this process of 'explicitation', learning is also observable in the 'proceduralization' that accompanies it, affecting the task as it gets automatized and allows for less conscious control. This holds for our singer when he can sing a difficult score by heart. In that case, an expert can expand his performance by adding further elements to it or further refining it. As the term 'neuroconstructivism' suggests, this theory entails that during learning, the underlying neural mechanism changes by developing a more complex, modular structure. We argued that this 'modularization' concurs largely with the 'kludge formation' that according to us tends to affect mechanisms. We emphasized another insight from neuroconstructivism, which is that as a result of learning, there are several representations available to an expert for the performance of a task and not just a single one. Important for the present context is the consequence that an expert can be distinguished by his capability of switching to different modes of processing when performing a particular task, which a novice cannot do.

Differences between processing modes are what inspired a set of 'dual-process theories', the topic of chapter II.3. These theories distinguish between an automatic and a controlled mode of processing, differing among other things with regard to the information load they can process, the involvement of conscious control and the role of explicit knowledge. It implies that an agent gradually acquires the capability of performing a particular task in both modes of processing, as automatic processing is a result of his experience. This can be problematic for an agent because automatic processing can yield results that are stereotypical, for example, and not always in line his performance in the other, controlled mode of processing. A singer performing Don Giovanni rather automatically may have difficulty avoiding a macho comportment, for example. We argued that such automatic processing is in itself beneficial for expert action, the important question being whether an agent can somehow control when his performance relies upon automatic processing or when it does not. Our discussion confirmed that some control is indeed available to the agent, pertaining to various aspects of his task performance. Regaining some control can be done by reducing the complexity of the information that is processed during the task, by changing its representation or by chunking it. Some self-regulation is possible, too, as when the agents somehow prepares for the conditions under which automatic instead of controlled processing would prevail. We argued that even such forms of self-regulation can lead to kludge formation and become integrated in the mechanisms responsible for automatic and controlled processing. So while admitting that task performance can rely upon different modes of processing, we rejected a strict separation of the two. Sculpting the space of his actions also implies that an agent improves upon his capability of regulating the different processing modes and the mechanisms involved that are responsible for his performances.

Chapter II.4 focused more specifically to a discussion of how external information becomes integrated in a mechanism responsible for an agent's expert action due to learning and development. Especially as humans often rely on representations that employ language or symbols when they are learning, practicing or adjusting a task performance, the question is whether kludge formation obtains. We argued that this is indeed the case. Adopting Barsalou's simulation theory, we explained expertise in terms of learners developing many 'simulators' that facilitate expert performance in a particular domain, like the domain of opera performance. A simulator consists of a complex, hierarchically structured network of component representations for a domain, which are stored in a distributed fashion across the brain and can be employed by different functions alike. Explicit representations and linguistic concepts can influence the formation, configuration and activation of these

representations. So when an expert action is performed, the agent in a sense 're-enacts' a previous experience or action, or he composes a novel one by employing his stored representations. An expert therefore has multiple advantages compared to a novice, as he can employ a sculpted space of actions and has expertise in its targeted use. Hence, learning a new opera role is easier for an expert than for a novice.

We continued this chapter by discussing the theory of extended cognition presented by Clark and Chalmers. This theory holds that some cognitive or behavioral tasks rely so much upon the properties of external tools or other objects, that we should even include these in the mechanistic explanation of such a task. We argued instead not to expand the responsible explanatory mechanism by including external objects in it, but to explain the amazing interactions with objects by way of the human capability of developing complex representations in which object properties are integrated. Such a representation can then affect the mechanism responsible for a task. In other words, we aligned the simulation theory and the theory of extended cognition by applying our methodological insights.

In sum, we showed in Part II that explaining how an agent can learn to perform an expert action like performing the role of Don Giovanni should indeed be done by using the explanatory ingredients prepared in Part I. Development and learning, so we concluded, can be understood in terms of changes that affect relevant mechanisms and representations. The result is a complex situation, as an expert can perform a certain task in more than just a single way, for example via automatic or controlled processing or by employing one or another task representation, which a novice cannot. It is thanks to the process of sculpting the space of actions that an expert finds himself in that comfortable position.

Part III is devoted to a more specific investigation of intentional action, applying the methodological resources prepared in Part I, and the insights regarding development and learning from Part II. Indeed, we demonstrated that the explanation of intentional action is comparable to the explanation of expert action. Surprising as this may seem, by navigating between conceptual analyses of the components of intentional action and their empirical study, we demonstrated that an agent can only consistently perform actions according to his intentions when he has been sculpting the space of his actions. Part III started with a chapter expounding the framework to be used when discussing action intentions. Next, consecutive chapters are devoted to these intentions, always navigating from philosophical analyses to a discussion of empirical studies.

In chapter III.1 we introduced the notion of 'sculpting the space of actions', which

was mentioned above. We clarified in section III.1.1 why it is valuable to explain a given task as a problem in finding an adequate option in a so-called search space. It particularly facilitates such an explanatory effort as it enables the integration of multiple determining factors by representing each factor as an extra dimension to this multi-dimensional space. Expertise, we argued, should accordingly be conceived as a sculpting process affecting in several ways this space as Frith has done in the context of a language-processing task. Extending his analysis, we distinguished both a long-term and short-term sculpting process, having stable and dynamic effects on several related tasks. For example, a novice with a less sculpted space will be far slower and less adequate in his responses, but also in his perception and understanding of novel situations because a sculpted space is being employed by several cognitive processes alike.

These insights concerning a sculpted space will be integrated with Pacherie's framework containing three different levels of intention: motor intentions are responsible for guiding ongoing motor movements, proximal intentions for anchoring an intention in a given situation and distal intentions are the long term decisions about future actions. In section III.1.2 we described this framework and showed how it understands and explains intentions by integrating philosophical analyses from Frankfurt, Bratman and others with cognitive neuroscientific insights from Jeannerod and others. The framework organizes the different levels of intentions in a hierarchical structure and together with their interactions these enable an agent to eventually realize in motor movements a complex action that he decided to do long before the appropriate situation occurred. In this 'intentional cascade' framework, action representations were again found to play a central role, inviting their integration in a multidimensional, sculpted space of actions. Having laid out these notions of the intentional cascade and the sculpted space of actions, we then started with the discussion of the lower level of the cascade: first in a section with a philosophical analysis, second in a section regarding empirical cognitive neuroscientific insights.

Section III.2.1 contained a philosophical analysis of why motor intentions are distinguished which guide ongoing body movements. The fact that actions are continuously, fast and minutely adjusted to internal and external conditions suggests that these motor intentions play a role by integrating information about action goals, movements and a changing environment. Frankfurt was found to underline that we can observe how an agent continually receives feedback about his action and adjusts it accordingly. Such adjustments occur because an action, being different from a mere reflex, must be taken to stand in a particular relation to the agent's overall identity as cognitive, affective and attitudinal processes have determined it, all contributing to

consistency in his actions – even at this level of motor intentions.

Section III.2.2 then discussed empirical studies of motor intentions from Jeannerod and others yielding results that suggest how in fact such determination and guiding of an ongoing action is implemented. A motor intention guiding an action is constituted by a motor representation in a non-conceptual format that integrates, promptly and without conscious control, not just information concerning muscular movements but also information concerning the environment and the affordances for action that it contains. Experience was found to influence these representations in several ways, sculpting the space of an agent's actions.

Section III.2.2 continued with reference to De Groot's seminal studies of experiments with chess masters, elucidating how expertise affected the representations involved in their expert actions. Sculpting the space of their actions, they were found to assemble a very large number of increasingly complex and – hierarchically – structured representations, facilitating simultaneously their expert perception, decision-making and actions in complex situations. Interested in the representational redescription involved, we discussed the 'template theory' developed by Gobet and Simon which explains why experts are not only capable of handling complex situations but also flexible in doing so: their representations consist not just of complete chunks of information but also of complex templates with free slots that remain open for variable information. We pointed out how corresponding to these changes in representation, two neural processes obtain during learning of expert action, affecting subserving mechanisms. At first, expertise implies an increasing efficiency of neural activations during task performance, second, co-activations obtain which are due to other neural representations or processes related to the task at hand. Besides, it was found that specific neural areas or even single neurons can represent specific components of these motor representations, which are employed not just for motor actions but also for other tasks. We concluded that learning does indeed lead to the generative entrenchment both of particular components of the mechanisms responsible for the guidance of an expert's actions and of the specific motor representations involved.

In section III.3.1 we offered a philosophical analysis of proximal intentions. These fulfill fulfill a mediating role between the motor intentions and the distal intentions by specifying an action intention in a motor intention format, even though the action intention is initially made in a conceptual format long before an appropriate situation presents itself. So an expert singer somehow has to anchor his practiced interpretation of Don Giovanni's arias and behavior in a situation with specific stage props, ongoing directions of the conductor, other singers and so on. We applied to this the explanatory tool we introduced earlier, considering such anchoring as the singer's

further sculpting his stably sculpted space of actions. Proximal intentions' mediating role is particularly evident when an agent blocks a habitual action in an exceptional situation or when it is overridden by another – conflicting - distal intention, according to Bratman. In such cases, constraints are derived from the more comprehensive web of intentions and action plans that an agent typically has, about which more later. Together, these contribute to the consistency of his actions that is even visible in expert motor actions, in defiance of the paradox of action.

In III.3.2 we investigated empirical studies regarding the implementation of proximal intentions and argued that its intermediating role likely involves not just one but two distinguishable processes. We scrutinized the model of Norman & Shallice and colleagues which enabled us to explain both the habitual nature of complex actions by an expert and the potential modification or blocking of a habitual action. According to this model, large knowledge structures or action representations play a central role. With the interaction between a 'contention scheduling' process involved in composing the representation for a habitual action, and a 'supervisory attentional system' that can modulate or intervene in that process, we succeeded in explaining various properties of a proximal intention. According to this model, action representations are composed of many loose components that are put together in a hierarchical organization. This assembly of an action representation depends on the interactive activations with which components are related to each other and with other features like environmental triggers or goal conditions, as a result of development and learning.

We continued section III.3.2 by explaining how a sculpted space of actions is characterized among others things by specific interactive activation patterns. Effects of these can be observed when the opera singer manages to switch quite effortlessly during a rehearsal from Don Giovanni's seduction scene to performing Saint François's dialogue with the birds. Upon the activation of a particular action component – for instance by hearing the introduction of an aria - associated component representations are activated pertaining to other aspects of the performance. Other actions are facilitated, different anticipations regarding the environment arise and other constraints that depend upon his distal intentions are activated, too. We explained the importance of preparation and practice of such complex actions helps, as it increases the interactive activations between action components and can modulate these in a targeted way. Applying our notion of how development can involve the formation of kludge within a responsible mechanism, we explained how an expert is able to anchor and specify his intentions so fast, flexibly and consistently in contrast to a novice. We finally touched upon the neural

implementation of proximal intentions before turning to distal intentions, posited at the highest level of the intentional cascade.

Motor intentions and proximal intentions were shown to play indispensable roles in the performance of intentional action, doing so with relative autonomy. Yet they were also found to be only indirectly or partly related to the distal intentions, even though the latter are usually considered to be genuine intentions. Moreover, within the intentional cascade, not just top-down but also bottom-up influences are at stake, suggesting that distal intentions are themselves also influenced by the contents of an agent's prominent motor intentions, notwithstanding their different representational formats. Addressing these and other issues, chapter III.4 offered an extensive account of the roles of distal intentions, their implementation in the form of imagination or narrative simulation, the socio-cultural nature of schemas involved in such simulations and hypotheses about their neural implementation.

Section III.4.1 started by discussing Bratman's philosophical account of distal intentions, arguing that they play an important role in the complex task of coordinating and organizing an agent's actions. Without distal intentions constraining his space of actions, an agent likely engages in counterproductive actions, is incapable of realizing complex and temporally extended actions, and must keep cognitive resources free for continuously forming his intentions. We argued that expert action would be impossible under those conditions. This also implies that these distal intentions should not be reconsidered or changed lightly, but provide stable structure to an agent's sculpted space of actions. We concluded, however, that this account unfortunately has little to say about how an agent can represent the complex web of all of his intentions, which would enable him to organize and coordinate his actions and intentions comprehensively.

For that reason, we continued the philosophical analysis of section III.4.1 by proposing to remedy shortcomings of this account of distal intentions by enlarging it with Ricoeur's theory of narrative configuration of action. This theory contends that agents always engage in narrative, making action configurations that extend beyond the contents of single distal intentions to three further hierarchical levels: first, the level of socially shared practices, second, the level of plans regarding family life, professional life, and the like, and third, the comprehensive level of the unity of a life. By configuring and reconfiguring his narrative, an agent integrates heterogeneous ingredients like action components, goals, values, and temporal structure, but also environmental conditions and chance. Important to note is that this complex task is influenced by configurations or representations that a culture or tradition provides, even if an agent inevitably deviates from such examples. By way of his narrative, an

agent can not only describe his – past, present and future - actions and intentions, but also explain and perhaps adjust them and thus develop his identity as an agent, for himself and others alike. We argued that through narrative, an agent at least has the resources to plan and coordinate but also to evaluate and weigh his distal intentions in a way that Bratman nor Pacherie's intentional cascade were found to present.

In section III.4.2 we proposed to consider the implementation of distal intentions in cognitive and neural processes along the lines of a simulation theory, similar to the one discussed earlier in chapter II.4: action representations are stored not as a whole but as a hierarchically organized network of component representations throughout the brain, which are employed by different cognitive functions. Repeated employment strengthens the connections between components of a representation, causing the representation to become more deeply entrenched and more likely to influence future tasks. The simulation theory presented by Schacter and others can explain distal intentions and narrative in terms of 'constructive memory' processes. In doing so, the theory integrates both representations and mechanisms and confirms the notion of a sculpted space of actions.

We continued section III.4.2. by discussing options for the neural implementation of narrative, recognizing its comprehensive task in coordinating, organizing and evaluating the agent's actions and intentions. We argued that the default-mode-network discovered by Raichle and others is a candidate, as it was found to be involved in maintaining and evaluating information and to be activated during interpretive and predictive tasks, including those that are self-referential in nature. Doing all this, it plays an additional and important role in sculpting the space of actions of an agent, enabling him to become an expert who consistently and coherently performs actions that comply with his intentions to the surprise of observers.

What have we reached with these navigations and why did they have to be so extended? At the outset of this section, we pointed out how observations, conceptualizations, investigations and explanations cohere intimately with each other, making the task of explaining human action complex. We intended this study to contribute to the necessary integration of intentions and mechanisms for the explanation of human action. We have argued why researchers should integrate insights about the representation of information with the mechanistic explanation of a task, considering representation as another causal factor. Furthermore, we have expanded the theory of mechanistic explanation by presenting four potential modifications of an explanatory mechanism, to be used for the explanation of development and learning. Applying this explanatory method, we have proposed to

explain the stable results of such dynamic processes as effects of kludge formation within a mechanism responsible for a given task. In addition, it was emphasized that both kludges and the associated representations can become generatively entrenched in the mechanism, giving rise to and taking part in subsequent developments, thus engendering a snow-ball effect. These insights regarding methodology and development were then applied to and helpful for explaining how actions are dependent upon different kinds, or levels, of intentions. We have introduced the concept of 'sculpting the space of actions', which enabled a comprehensive account of intentional action and the effects of expertise on it. Finally, we argued why the narrative configuration of action should be added to the intentional cascade, which also contributes to an expert's sculpted space of actions.

It is not uncommon to complain that philosophers are better at raising new questions than at answering current ones. We hope to have shown in a modest sense that these in fact cohere intimately: by asking attention for issues or relations that have been neglected somewhat, existing problems often appear in a new light. Philosophical contributions can in that sense make useful contributions to the complex interdisciplinary investigation of human action, inviting as it does adequate questions and answers from as fields as diverse as philosophy, cognitive neuroscience, social science, robotics, computational and animal studies, and more. However, scientific projects rarely lead to the development of genuine interdisciplinary and comprehensive accounts but focus instead on further clarification of more specific features or elements. This can easily lead to rash and simplified accounts, which certainly has happened with respect to human action. The paradox of expert action, with which we started this section, is a point in case, as it depends in part on a misunderstanding of what lies behind automatic expert performance. Similar examples can be found in the debates about free will, a topic that we had to leave to another time. Yet the observation of expert action as a result of an agent's long-term deliberately sculpting the space of his actions should make us pause about the rejection of the importance of free will for human action in general. In contrast to many who decry human intentional and voluntary action as being inexistent, impossible, outclassed or otherwise absent, this dissertation can also be read as an argument that intentional action is in fact possible, yet reliant on mechanisms and intentions that are more complex than often assumed. Indeed, our argument may even be taken as supporting the importance of not only musical but also moral education and practice: the admiration we feel for an expert opera singer or a moral hero is more than justified and should inspire us to likewise sculpt our space of actions.

**Figure I. Representation of an explanatory mechanism responsible for Φ-ing and some of its possible modifications, as discussed in Part I**



Simplified representations of the explanatory mechanism responsible for the task - or explanandum phenomenon - Φ-ing at two different moments in time, t1 and t2. Between those moments several kinds of mechanism modification have occurred at different levels, as was discussed in chapter I.5. Mechanism modifications represented here are: increased (bold) and decreased (striped) interactions between components; a new feedback relation (yellow) between 'q' and 'r'; new influence (blue) of component 't' on component 'r'; new interaction (blue) between Ψ-ing and Φ-ing, yielding an - indirect - influence of new input to Φ-ing. These mechanism modifications are together responsible for modifications of the cognitive and behavioral outputs of Φ-ing. Note that vertical red dotted lines represent constitutive relations between mechanism levels and that green dotted circles represent the context within which components figure.

# Figure II. Representation of an explanatory mechanism involved in development or learning as accounted for by neuroconstructivist theories, as discussed in Part II
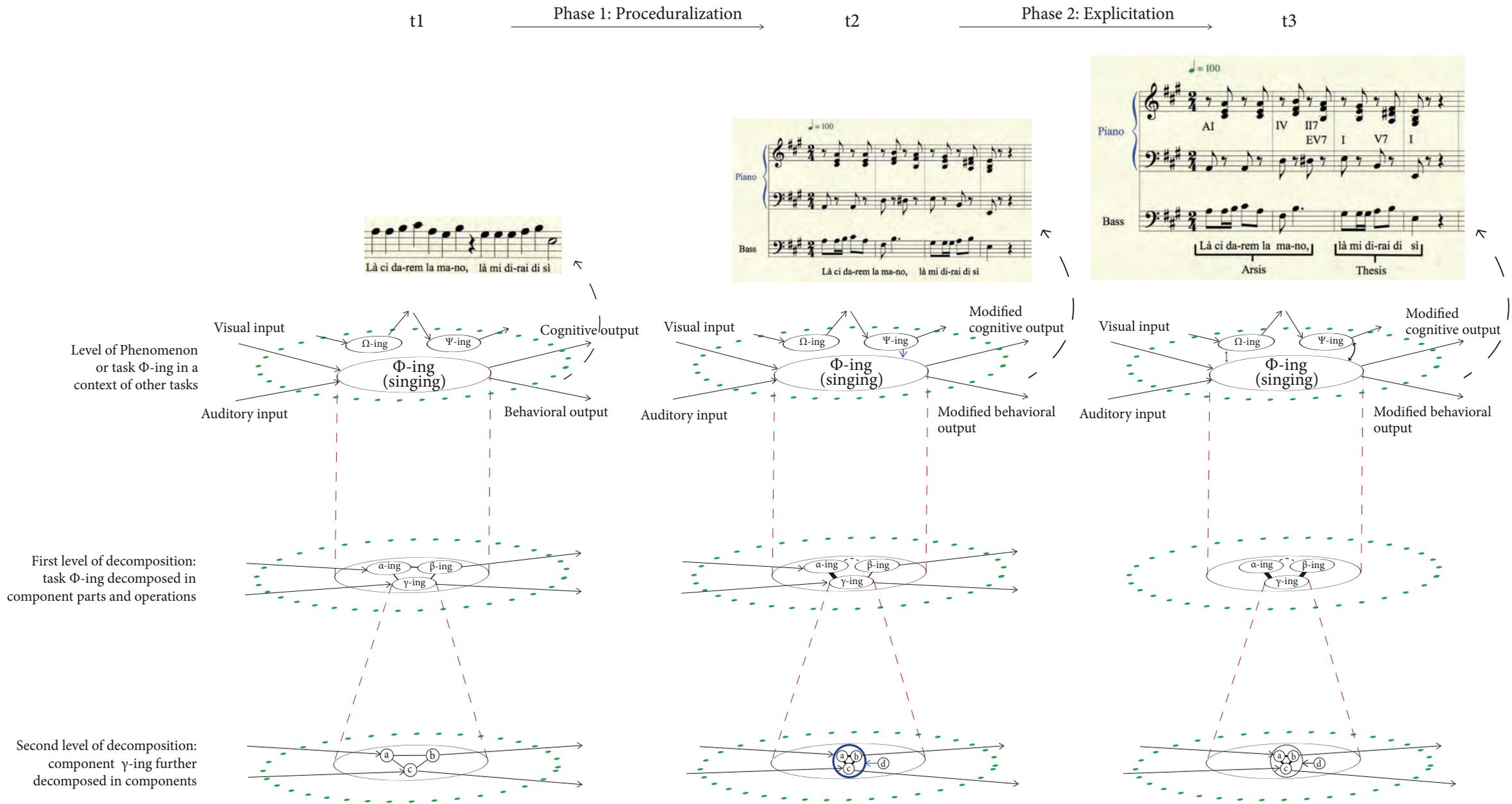


Simplified representations of the explanatory mechanism responsible for the task Φ-ing, in this case: singing. According to the neuroconstructivist theories of development and learning discussed in chapter II.2, two (partly overlapping) phases of learning can be distinguished. Mechanism modifications are involved in this (see figure I). The first phase of proceduralization is characterized by improved performance of Φ-ing, enabled by the modularization (blue circle) of some mechanism (sub-)components which enables their faster and stable processing, enabling increased (bold) and decreased (dotted) interactions between components at another level. This also facilitates additional interaction (blue) between tasks Φ-ing and Ψ-ing (e.g. acting). The second phase of explicitation is then characterized by increasing explicit control of Φ-ing, again facilitated by additional interactions (blue) between tasks. Due to representational redescriptions, multiple representations involved in the task (curved dotted arrows) are available to the agent, leading from an implicit and simple representation to more complex, hierarchically structured and explicit representations of the music piece. Note that vertical red dotted lines represent constitutive relations between mechanism levels and that green dotted circles represent the context within which components figure.

**Figure III. Representation of explanatory mechanisms responsible for Φ-ing in a novice and an expert in a particular situation with their distinct sculpted spaces of actions, as discussed in Part III**
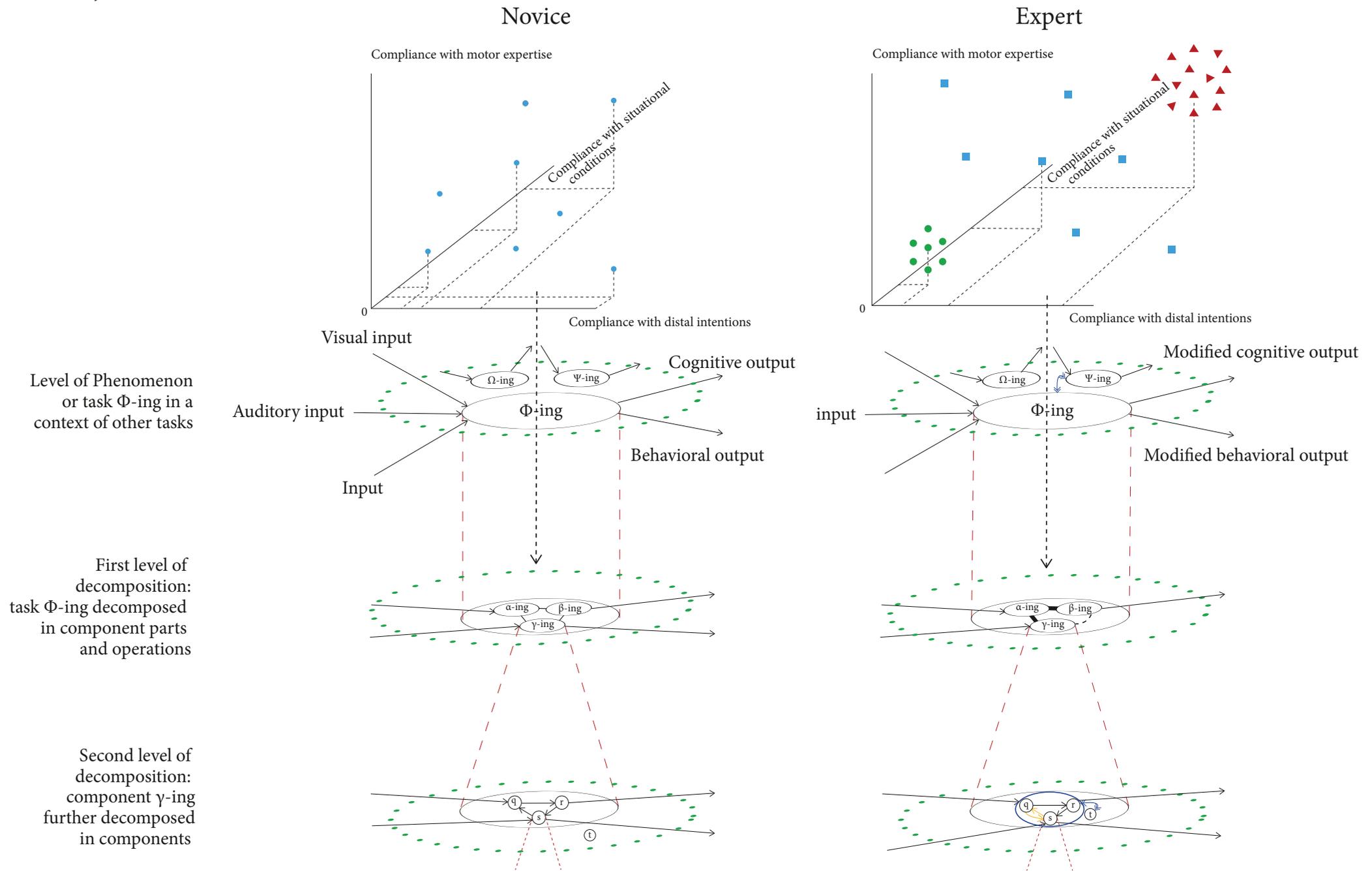


Simplified representation of the explanatory mechanisms responsible for the task - or explanandum phenomenon - Φ-ing in a novice and an expert respectively, in this case the task of determining an action in a particular situation, as was discussed in chapter III.4. With his increased expertise, the expert's responsible mechanism has become modified (see figures II and III). Corresponding with that process, the space of actions from which an action will now be determined by the interacting (sub-)components of Φ-ing has become sculpted in the expert, which is not the case for the novice's space of actions. Red dots represent preferred action options that maximally comply with his motor expertise (vertical axe), with his distal intentions (horizontal axe) and with the situational conditions (diagonal axe). Green dots represent suppressed action options that comply only minimally. Blue dots refer to action actions that are neither preferred nor suppressed. The novice's space of actions contains only a few of those blue dots. Note that vertical red dotted lines represent constitutive relations between mechanism levels and that green dotted circles represent the context within which components figure.

# REFERENCES

Aarts, H., & Dijksterhuis, A. (2000). Habits as Knowledge Structures: Automaticity in Goal-Directed Behavior. *Journal of Personality and Social Psychology, 78*(1), pp. 53-63.

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14*(1), pp. 43-64.

Addis, D.R., Musicaro, R., Pan, L., et al. (2010). Episodic Simulation of Past and Future Events in Older Adults: Evidence From an Experimental Recombination Task. *Psychology and Aging, 25*(2), pp. 369-376.

Addis, D.R., Sacchetti, D.C., Ally, B.A., et al. (2009). Episodic simulation of future events is impaired in mild Alzheimer's disease. *Neuropsychologia, 47*(12), pp. 2660-2671.

Adi-Japha, E., Berberich-Artzi, J., & Libnawi, A. (2010). Cognitive Flexibility in Drawings of Bilingual Children. *Child Development, 81*(5), pp. 1356-1366.

Adolphs, R. (2009). The Social Brain: Neural Basis of Social Knowledge. *Annual Review of Psychology, 60*(1), pp. 693-716.

Aiktns, D., & Ray, W.J. (2001). Frontal lobe contributions to hypnotic susceptibility: A neuropsychological screening of executive functioning. *International Journal of Clinical and Experimental Hypnosis, 49*(4), pp. 320-329.

Aisbett, J., & Gibbon, G. (2001). A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence, 133*(1–2), pp. 189-232.

Aizawa, K. (2010). The coupling-constitution fallacy revisited. *Cognitive Systems Research, 11*(4), pp. 332-342.

Ambrose, S.H. (2001). Paleolithic Technology and Human Evolution. *Science, 291*(5509), pp. 1748-1753.

Ames, D.L., & Fiske, S.T. (2010). Cultural neuroscience. *Asian Journal of Social Psychology, 13*(2), pp. 72-82.

Anderson, J.R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), pp. 369-406.

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J.R., & Schooler, L.J. (1991). Reflections of the Environment in Memory. *Psychological Science, 2*(6), pp. 396-408.

Anderson, M.L. (2003). Embodied Cognition: A field guide. *Artificial Intelligence, 149*(1), pp. 91-130.

Anderson, M.L. (2007). Massive redeployment, exaptation, and the functional integration of cognitive operations. *Synthese, 159*(3), pp. 329-345.

Anderson, M.L. (2008). Are interactive specialization and massive redeployment compatible? *Behavioral and Brain Sciences, 31*(03), pp. 331-334.

Anderson, M.L. (2010). Neural re-use as a fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33*(04), pp. 245-266.

Anderson, M.L., Richardson, M.J., & Chemero, A. (2012). Eroding the Boundaries of Cognition: Implications of Embodiment. *Topics in Cognitive Science, 4*(4), pp. 717-730.

Andrews-Hanna, J.R., Snyder, A.Z., Vincent, J.L., et al. (2007). Disruption of large-scale brain systems in advanced aging. *Neuron, 56*(5), pp. 924-935.

Annas, J. (1977). How Basic Are Basic Actions? *Proceedings of the Aristotelian Society, 78*, pp. 195-213.

Arbib, M., & Bonaiuto, J. (2007). From grasping to complex imitation: mirror systems on the path to language. *Mind & Society, 7*(1), pp. 43-64.

Arbib, M.A. (1981). Perceptual Structures and Distributed Motor Control. In V.B. Brooks (Ed.), *Handbook of Physiology – The Nervous System II. Motor Control* Bethesda, MD: American Physiological Society, pp. 1449-1480.

Arbib, M.A. (2003). Rana computatrix to human language: Towards a computational neuroethology of language evolution. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences (Series A), 361*(1811), pp. 2345-2379.

Arbib, M.A. (2003). Schema theory. In M.A. Arbib (Ed.), *The handbook of brain theory and neural networks* Cambridge, M.A.: MIT Press.

Arbib, M.A. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences, 28*(2), pp. 105-124.

Arbib, M.A. (2011). From Mirror Neurons to Complex Imitation in the Evolution of Language and Tool Use. *Annual Review of Anthropology, 40*(1), pp. 257-273.

Arbib, M.A., Bonaiuto, J.B., Jacobs, S., et al. (2009). Tool use and the distalization of the end-effector. *Psychological Research-Psychologische Forschung, 73*(4), pp. 441-462.

Archibald, S., Mateer, C., & Kerns, K. (2001). Utilization Behavior: Clinical Manifestations and Neurological Mechanisms. *Neuropsychology Review, 11*(3), pp. 117-130.

Aristotle. (1972). *Aristotle's De partibus animalium I and De generatione animalium I  (with passages from II. 1-3)* (D.M. Balme, Trans.). Oxford: Clarendon Press.

Aristotle. (1984). *The complete works of Aristotle: the revised Oxford translation* (Edited by Barnes, J.). Princeton: Princeton University Press.

Arnett, J.J. (2008). The Neglected 95%: Why American Psychology Needs to Become Less American. . *American Psychologist, 63*(7), pp. 602-614.

Aron, A.R. (2011). From Reactive to Proactive and Selective Control: Developing a Richer Model for Stopping Inappropriate Responses. *Biological Psychiatry, 69*(12), pp. 55-68.

Ashby, F.G., & Ell, S.W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences, 5*(5), pp. 204-210.

Ashby, F.G., Turner, B.O., & Horvitz, J.C. (2010). Cortical and basal ganglia contributions to

habit learning and automaticity. *Trends in Cognitive Sciences, 14*(5), pp. 208-215.

Atance, C.M. (2008). Future Thinking in Young Children. *Current Directions in Psychological Science, 17*(4), pp. 295-298.

Atance, C.M., & O'Neill, D.K. (2001). Episodic future thinking. *Trends in Cognitive Sciences, 5*(12), pp. 533-539.

Atkinson, A.P., Thomas, M.S.C., & Cleeremans, A. (2000). Consciousness: mapping the theoretical landscape. *Trends in Cognitive Sciences, 4*(10), pp. 372-382.

Audi, R. (1991). Responsible Action and Virtuous Character. *Ethics, 101*(2), pp. 304-321.

Baddeley, A.D. (1976). *The psychology of memory*. New York, NY: Harper & Row.

Baddeley, A.D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), pp. 829-839.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences, 12*(5), pp. 193-200.

Baldwin, D.A., & Baird, J.A. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences, 5*(4), pp. 171-178.

Baldwin, D.A., Andersson, A., Saffran, J., et al. (2008). Segmenting dynamic human action via statistical structure. *Cognition, 106*(3), pp. 1382-1407.

Baliki, M.N., Geha, P.Y., Apkarian, A.V., et al. (2008). Beyond Feeling: Chronic Pain Hurts the Brain, Disrupting the Default-Mode Network Dynamics. *J. Neurosci., 28*(6), pp. 1398-1403.

Bar, M. (2009). The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), pp. 1235-1243.

Barbey, A.K., Krueger, F., & Grafman, J. (2009). Structured event complexes in the medial prefrontal cortex support counterfactual representations for future planning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), pp. 1291-1300.

Bargh, J.A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* New York: Guilford Press, pp. 361-382.

Bargh, J.A., & Ferguson, M.J. (2000). Beyond behaviorism: on the automaticity of higher mental processes. *Psychological Bulletin, 126*(6), pp. 925.

Bargh, J.A., Gollwitzer, P.M., Lee-Chai, A., et al. (2001). The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals. *Journal of Personality & Social Psychology, 81*(6), pp. 1014-1027.

Barker, F.G. (1995). Phineas among the phrenologists: the American crowbar case and nineteenth-century theories of cerebral localization. *J. of Neurosurgery, 82*(4), pp. 672-682.

Barresi, J., & Moore, C. (1996). Intentional relations and social understanding. *Behavioral and brain sciences, 19*(1), pp. 107-121.

Barrett, H.C., & Kurzban, R. (2006). Modularity in Cognition: Framing the Debate.

*Psychological Review, 113*(3), pp. 628-647.

Barsalou, L.W. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes, 28*(1), pp. 61-80.

Barsalou, L.W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences, 22*(04), pp. 637-660.

Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), pp. 577-609.

Barsalou, L.W. (2002). Being there conceptually: Simulating categories in preparation for situated action *Representation, memory, and development: Essays in honor of Jean Mandler.*: Mahwah, NJ, US: Lawrence Erlbaum, pp. 1-15.

Barsalou, L.W. (2003). Situated simulation in the human conceptual system. *Language and Cognitive Processes, 18*(5-6), pp. 513-562.

Barsalou, L.W. (2005). Continuity of the conceptual system across species. *Trends in Cognitive Sciences, 9*(7), pp. 309.

Barsalou, L.W. (2008). Grounded Cognition. *Annual Review of Psychology, 59*(1), pp. 617-645.

Barsalou, L.W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), pp. 1281-1289.

Barsalou, L.W. (2010). Grounded Cognition: Past, Present, and Future. *Topics in Cognitive Science, 2*(4), pp. 716-724.

Barsalou, L.W. (2011). Integrating Bayesian analysis and mechanistic theories in grounded cognition. *Behavioral and Brain Sciences, 34*(4), pp. 191-192.

Barsalou, L.W., Barbey, A.K., Simmons, W.K., et al. (2005). Embodiment in Religious Knowledge. *Journal of Cognition and Culture, 5*, pp. 14-57.

Barsalou, L.W., Cohen, H., & Lefebvre, C. (2005). Situated Conceptualization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* Oxford: Elsevier, pp. 619-650.

Barsalou, L.W., Simmons, W.K., Barbey, A.K., et al. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*(2), pp. 84.

Bartlett, F.C. (1995 [1932]). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Bassett, D.S., & Gazzaniga, M.S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences, 15*(5), pp. 200-209.

Bassett, D.S., Wymbs, N.F., Porter, M.A., et al. (2011). Dynamic reconfiguration of human brain networks during learning. *PNAS, 108*(18), pp. 7641-7646.

Bateson, P. (2008). A good approach to neural and behavioural development but would be even better if set in a broader context. *Behavioral and Brain Sciences, 31*(03), pp. 334-335.

Baumeister, R.F., Crescioni, A.W., & Alquist, J.L. (2011). Free will as advanced action control

for human social life and culture. *Neuroethics, 4*(1), pp. 1-11.

Beatty, J. (1997). Why Do Biologists Argue like They Do? *Philosophy of Science, 64*, pp. S432-S443.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines, 4*(1), pp. 1-25.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22*(3), pp. 295-317.

Bechtel, W. (2001). The Compatibility of Complex Systems and Reduction: A Case Analysis of Memory Research. *Minds and Machines, 11*(4), pp. 483-502.

Bechtel, W. (2001a). Decomposing and Localizing Vision: an Exemplar for Cognitive Neuroscience. In W. Bechtel, P. Mandik, J. Mundale & R. Stufflebeam (Eds.), *Philosophy and the Neurosciences. A Reader.* Malden, MA: Blackwell, pp. 225-249.

Bechtel, W. (2001b). Linking cognition and brain: The cognitive neuroscience of language. In W. Bechtel, P. Mandik, J. Mundale & R. Stufflebeam (Eds.), *Philosophy and the Neurosciences. A Reader.* Malden, MA: Blackwell, pp. 152-171.

Bechtel, W. (2002). Decomposing the Mind-Brain: A Long-Term Pursuit. *Brain and Mind, V3*(2), pp. 229-242.

Bechtel, W. (2008). *Mental mechanisms. Philosophical Perspectives on Cognitive Neuroscience.* New York: Lawrence Erlbaum Associates.

Bechtel, W. (2009). Explanation: Mechanism, Modularity, and Situated Cognition. In P. Robbines & M. Aydede (Eds.), *Cambridge handbook of situated cognition* Cambridge: Cambridge University Press, pp. 155-170.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology, 22*(5), pp. 543 - 564.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), pp. 421-441.

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies In History and Philosophy of Science Part A, 41*(3), pp. 321-333.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, pp. 175-207.

Bechtel, W., & Richardson, R.C. (1993). *Discovering complexity: decomposition and localization as strategies in scientific research.* Princeton, NJ: Princeton University Press.

Bedwell, J.S., Gallagher, S., Whitten, S.N., et al. (2011). Linguistic correlates of self in deceptive oral autobiographical narratives. *Consciousness and Cognition, 20*(3), pp. 547-555.

Beilock, S.L. (2009). Grounding cognition in action: expertise, comprehension, and judgment.

In M. Raab, J.G. Johnson & H.R. Heekeren (Eds.), *Progress in Brain Research* Vol. Volume 174: Elsevier, pp. 3-11.

Bellenkes, A.H., Wickens, C.D., & Kramer, A.F. (1997). Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicine, 68*(7), pp. 569-579.

Benner, P. (2004). Using the Dreyfus Model of Skill Acquisition to Describe and Interpret Skill Acquisition and Clinical Judgment in Nursing Practice and Education. *Bulletin of Science Technology Society, 24*(3), pp. 188-199.

Bennett, M.R. (2007). Neuroscience and philosophy. In M. Bennett, D.C. Dennett, P.M.S. Hacker & J.R. Searle (Eds.), *Neuroscience and philosophy brain, mind, and language* New York: Columbia University Press.

Bennett, M.R., & Hacker, P.M.S. (2003). *Philosophical foundations of neuroscience.* Malden, MA: Blackwell.

Bennett, M.R., & Hacker, P.M.S. (2007). The conceptual presuppositions of cognitive neuroscience. A reply to critics. In M. Bennett, D.C. Dennett, P.M.S. Hacker & J.R. Searle (Eds.), *Neuroscience and philosophy. Brain, mind, and language.* New York: Columbia University Press.

Bermúdez, J., & Cahen, A. (2012). Nonconceptual mental content. *The Stanford Encyclopedia of Philosophy* Spring 2012. Retrieved July 22, 2012, from http://plato.stanford.edu/entries/content-nonconceptual/

Bermúdez, J.L. (1995). Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States. *Mind & Language, 10*(4), pp. 333-369.

Berntson, G.G., & Cacioppo, J.T. (2008). A contemporary perspective on multilevel analyses and social neuroscience. In F.S. Kessel, P.L. Rosenfield & N.B. Anderson (Eds.), *Interdisciplinary research: case studies from health and social science* New York: Oxford University Press.

Bialystok, E., Shenfield, T., & Codd, J. (2000). Languages, scripts, and the environment: Factors in developing concepts of print. *Developmental Psychology, 36*(1), pp. 66-76.

Bickle, J. (2003). Empirical evidence for a narrative concept of self. *Narrative and consciousness: Literature, psychology, and the brain*, pp. 195-208.

Bilalić, M., McLeod, P., & Gobet, F. (2009). Specialization Effect and Its Influence on Memory and Problem Solving in Expert Chess Players. *Cognitive Science, 33*(6), pp. 1117-1143.

Bischof-Köhler, D. (1985). Zur Phylogenese menschlicher Motivation. In L.H. Eckensberg & E.D. Lantermann (Eds.), *Emotion und Reflexivität* München: Urban und Schwarzenberg, pp. 3-47.

Blair, M.R., Watson, M.R., Walshe, R.C., et al. (2009). Extremely Selective Attention: Eye-Tracking Studies of the Dynamic Allocation of Attention to Stimulus Features in

Categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(5), pp. 1196-1206.

Bliss, T.V.P., & Collingridge, G.L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature, 361*(6407), pp. 31-39.

Bloch, M. (1991). Language, Anthropology and Cognitive Science. *Man, 26*(2), pp. 183-198.

Boden, M.A. (2004). *The creative mind. Myths and mechanisms.* London: Routledge.

Boncoddo, R., Dixon, J.A., & Kelley, E. (2010). The emergence of a novel representation from action: evidence from preschoolers. *Developmental Science, 13*(2), pp. 370-377.

Bonifazi, S., Farne, A., Rinaldesi, L., et al. (2007). Dynamic size-change of peri-hand space through tool-use: Spatial extension or shift of the multi-sensory area. *Journal of Neuropsychology, 1*, pp. 101-114.

Borrett, D., Kelly, S., & Kwan, H. (2000). Phenomenology, dynamical neural networks and brain function. *Philosophical Psychology, 13*(2), pp. 213 - 228.

Botvinick, M., & Plaut, D.C. (2004). Doing Without Schema Hierarchies: A Recurrent Connectionist Approach to Normal and Impaired Routine Sequential Action. *Psychological Review, 111*(2), pp. 395-429.

Botvinick, M.M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences, 12*(5), pp. 201-208.

Botvinick, M.M., Braver, T.S., Barch, D.M., et al. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review., 108*(3), pp. 624-652.

Botvinick, M.M., Niv, Y., & Barto, A.C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition, 113*(3), pp. 262-280.

Boyd, L.A., Edwards, J.D., Siengsukon, C.S., et al. (2009). Motor sequence chunking is impaired by basal ganglia stroke. *Neurobiology of Learning and Memory, 92*(1), pp. 35-44.

Boysen, S.T., Berntson, G.G., Hannan, M.B., et al. (1996). Quantity-based interference and symbolic representations in chimpanzees (Pan troglodytes). *Journal of Experimental Psychology / Animal Behavior Processes, 22*(1), pp. 76.

Brass, M., & Haggard, P. (2008). The What, When, Whether Model of Intentional Action. *Neuroscientist, 14*(4), pp. 319-325.

Bratman, M.E. (1981). Intention and Means-End Reasoning. *The Philosophical Review, 90*(2), pp. 252-265.

Bratman, M.E. (1984). Two Faces of Intention. *The Philosophical Review, 93*(3), pp. 375-405.

Bratman, M.E. (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

Bratman, M.E. (1992a). Planning and the stability of intention. *Minds and Machines, 2*(1), pp. 1-16.

Bratman, M.E. (1992b). Shared Cooperative Activity. *The Philosophical Review, 101*(2), pp. 327-341.

Bratman, M.E. (1999a). *Faces of intention: Selected essays on intention and agency.* Cambridge: Cambridge University Press.

Bratman, M.E. (1999b). Practical Reasoning and Acceptance in a Context *Faces of intention: selected essays on intention and agency* Cambridge: Cambridge University Press.

Bratman, M.E. (2002). Hierarchy, circularity, and double reduction. In S. Buss & L. Overton (Eds.), *Contours of Agency: Essays on Themes from Harry Frankfurt* Cambridge, MA: MIT Press, pp. 65-85.

Bratman, M.E. (2003). Autonomy and hierarchy. *Social Philosophy and Policy, 20*(02), pp. 156-176.

Bratman, M.E. (2006). *Planning agency and consciousness.* Paper presented at the *NWO symposium on Consciousness* on June 28, 2006, Utrecht.

Bratman, M.E. (2006). *Structures of Agency: Essays.* New York: Oxford University Press.

Bratman, M.E. (2007). Anchors for Deliberation. In C. Lumer & S. Nannini (Eds.), *Intentionality, Deliberation and Autonomy: The Action-Theoretic Basis of Practical Philosophy* Hampshire: Ashgate, pp. 187-205.

Bratman, M.E. (2009). Intention, Practical Rationality, and Self-Governance. *Ethics, 119*(3), pp. 411-443.

Bratman, M.E. (2009). Reflections on the philosophy of action. In J. Aguilar & A.A. Buckareff (Eds.), *Philosophy of Action: 5 Questions* Copenhagen: Automatic Press/VIP.

Bratman, M.E., Israel, D.J., & Pollack, M.E. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4*(3), pp. 349-355.

Broadbent, A. (2011). Inferring causation in epidemiology: mechanisms, black boxes, and contrasts. In P.I. McKay, F. Russo & J. Williamson (Eds.), *Causality in the Sciences* Oxford: Oxford University Press.

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1-3), pp. 139-159.

Broyd, S.J., Demanuele, C., Debener, S., et al. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews, 33*(3), pp. 279-296.

Bruce, L., Farrow, D., & Raynor, A. (2012). How specific is domain specificity: Does it extend across playing position? *Journal of Science and Medicine in Sport, 15*(4), pp. 361-367.

Bruner, J. (1991). The Narrative Construction of Reality. *Critical Inquiry, 18*(1), pp. 1-21.

Bruner, J., & Kalmar, D.A. (1998). Narrative and metanarrative in the construction of self. In M. Ferrari & R.J. Sternberg (Eds.), *Self-awareness: Its nature and development* New York: Guilford Press, pp. 308-331.

Buckner, R.L., & Carroll, D.C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), pp. 49-57.

Buekens, F. (2001). Essential indexiality and the irreducibility of phenomenal concepts. *Communication and Cognition, 34*(1-2), pp. 75-97.

Buekens, F., Vanmechelen, X., & Maessen, K. (2001). Indexicaliteit en dynamische intenties. *Algemeen Nederlands Tijdschrift voor Wijsbegeerte, 93*(3), pp. 165-180.

Buller, D.J., & Hardcastle, V.G. (2000). Evolutionary Psychology, Meet Developmental Neurobiology: Against Promiscuous Modularity. *Brain and Mind, 1*(3), pp. 307-325.

Buonomano, D.V., & Merzenich, M.M. (1998). Cortical Plasticity: From Synapses to Maps. *Annual Review of Neuroscience, 21*(1), pp. 149-186.

Buss, S. (2002). The True, the Good, and the Lovable: Frankfurt's Avoidance of Objectivity. In S. Buss & L. Overton (Eds.), *Contours of agency: essays on themes from Harry Frankfurt* Cambridge, MA: MIT Press.

Byrne, R.W., & Bates, L.A. (2007). Sociality, Evolution and Cognition. *Current Biology, 17*(16), pp. R714-R723.

Byrne, R.W., & Russon, A.E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences, 21*(05), pp. 667-684.

Callebaut, W., & Rasskin-Gutman, D. (Eds.). (2005). *Modularity. Understanding the Development and Evolution of Natural Complex Systems*. Cambridge, MA: MIT Press.

Calvo-Merino, B., Ehrenberg, S., Leung, D., et al. (2010). Experts see it all: configural effects in action observation. *Psychological Research-Psychologische Forschung, 74*(4), pp. 400-406.

Campitelli, G., & Gobet, F. (2010). Herbert Simon's decision-making approach: Investigation of cognitive processes in experts. *Review of General Psychology, 14*(4), pp. 354.

Campos, R., & Sotillo, M. (2008). Constructing minds: The development of mindreading abilities in typical and atypical trajectories. *Behavioral and Brain Sciences, 31*(03), pp. 336-337.

Cannizzaro, M., & Coelho, C. (2012). Analysis of Narrative Discourse Structure as an Ecologically Relevant Measure of Executive Function in Adults. *Journal of Psycholinguistic Research*, pp. 1-23.

Caracciolo, M. (2012). Narrative, meaning, interpretation: an enactivist approach. *Phenomenology and the Cognitive Sciences, 11*(3), pp. 367-384.

Carr, D. (1991). *Time, narrative, and history*. Bloomington, IN: Indiana University Press.

Casebeer, W.D., & Churchland, P.S. (2003). The Neural Mechanisms of Moral Cognition: A Multiple-Aspect Approach to Moral Judgment and Decision-Making. *Biology and Philosophy, 18*(1), pp. 169-194.

Casile, A., Caggiano, V., & Ferrari, P.F. (2011). The Mirror Neuron System: A Fresh View. *The Neuroscientist, 17*(5), pp. 524-538.

Catmur, C., Mars, R.B., Rushworth, M.F., et al. (2011). Making Mirrors: Premotor Cortex Stimulation Enhances Mirror and Counter-mirror Motor Facilitation. *Journal of Cognitive*

*Neuroscience, 23*(9), pp. 2352-2362.

Cattaneo, L., Caruana, F., Jezzini, A., et al. (2009). Representation of Goal and Movements without Overt Motor Behavior in the Human Motor Cortex: A Transcranial Magnetic Stimulation Study. *J. Neurosci., 29*(36), pp. 11134-11138.

Cattaneo, L., & Rizzolatti, G. (2009). The Mirror Neuron System. *Arch Neurol, 66*(5), pp. 557-560.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology.* New York: Guilford Press.

Chalmers, D. (2007). The hard problem of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* Malden, MA: Blackwell.

Chalmers, D.J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions. .* Cambridge, MA: MIT Press, pp.17-39.

Chambers, D., & Reisberg, D. (1985). Can Mental Images Be Ambiguous? *Journal of Experimental Psychology: Human Perception and Performance, 11*(3), pp. 317-328.

Changeux, J.P., & Ricoeur, P. (2000). *What makes us think? A neuroscientist and a philosopher argue about ethics, human nature, and the brain* (M.B. DeBevoise, Trans.). Princeton: Princeton University Press.

Charles, D. (1984). *Aristotle's philosophy of action.* Ithaca, New York: Cornell U.P.

Charles, D. (2000). *Aristotle on meaning and essence.* Oxford: Clarendon Press.

Charness, N., & Tuffiash, M. (2008). The Role of Expertise Research and Human Factors in Capturing, Explaining, and Producing Superior Performance. *Human Factors, 50*(3), pp. 427-432.

Chase, W.G., & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), pp. 55-81.

Chassy, P., & Gobet, F. (2011). A Hypothesis About the Biological Basis of Expert Intuition. *Review of General Psychology, 15*(3), pp. 198-212.

Chaturvedi, S.K., & Bhugra, D. (2007). The concept of neurosis in a cross-cultural perspective. *Current Opinion in Psychiatry, 20*(1), pp. 47-51.

Chella, A., Gaglio, S., & Pirrone, R. (2001). Conceptual representations of actions for autonomous robots. *Robotics and Autonomous Systems, 34*(4), pp. 251-263.

Cheon, B.K., Im, D.-m., Harada, T., et al. (2011). Cultural influences on neural basis of intergroup empathy. *NeuroImage, 57*(2), pp. 642-650.

Cherry, K., & Goerner, E.A. (2006). Does Aristotle's Polis Exist By Nature? *History of Political Thought, 27*, pp. 563-585.

Chiao, J.Y., Cheon, B.K., Pornpattananangkul, N., et al. (2013). Cultural Neuroscience: Progress and Promise. *Psychological Inquiry, 24*(1), pp. 1-19.

Chiao, J.Y., Harada, T., Komeda, H., et al. (2010). Dynamic Cultural Influences on Neural

Representations of the Self. *Journal of Cognitive Neuroscience, 22*(1), pp. 1-11.

Choudhury, S., & Gold, I. (2011). 'Special Issue on Cultural Neuroscience', Social Cognitive Affective Neuroscience. *BioSocieties, 6*(2), pp. 271-275.

Christensen, W., & Sutton, J. (2012). Reflections on Emotions, Imagination, and Moral Reasoning Toward an Integrated, Multidisciplinary Approach to Moral Cognition. In R. Langdon & C. Mackenzie (Eds.), *Emotions, Imagination, and Moral Reasoning* New York: Psychology Press, pp. 327-347.

Christian, B.M., Miles, L.K., Fung, F.H.K., et al. (2013). The shape of things to come: Exploring goal-directed prospection. *Consciousness and Cognition, 22*(2), pp. 471-478.

Chu, D. (2008). Criteria for conceptual and operational notions of complexity. *Artificial Life, 14*(3), pp. 313-323.

Churchland, P. (1988). Folk Psychology And The Explanation Of Human Behaviour. *Aristotelian Society, SUPP 62*, pp. 209-221.

Churchland, P.M. (1995). *The engine of reason, the seat of the soul : a philosophical journey into the brain*. Cambridge, MA: MIT Press.

Churchland, P.M. (1998). Toward a Cognitive Neurobiology of the Moral Virtues. *Topoi, 17*(2), pp. 83-96.

Churchland, P.S. (1986). *Neurophilosophy: toward a unified science of the mind/brain.* . Cambridge, MA: MIT Press

Churchland, P.S. (2002). How do brains represent? *Brain-Wise: Studies in Neurophilosophy* Cambridge MA: MIT Press, pp. 273-319.

Churchland, P.S. (2005). A neurophilosophical slant on consciousness research *Progress in Brain Research* Vol. 149, pp. 285-293.

Churchland, P.S., & Sejnowski, T.J. (1988). Perspectives on Cognitive Neuroscience. *Science, 242*(4879), pp. 741-745.

Cisek, P. (2006). Integrated Neural Processes for Defining Potential Actions and Deciding between Them: A Computational Model. *J. of Neuroscience, 26*(38), pp. 9761-9770.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philos. Trans. of the Royal Soc. B: Biological Sci., 362*(1485), pp. 1585-1599.

Cisek, P., & Kalaska, J.F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience, 33*(1), pp. 269-298.

Clark, A. (1987). The Kludge in the Machine. *Mind & Language, 2*(4), pp. 277-300.

Clark, A. (1997). *Being There. Putting Brain, Body, and World together again*. Cambridge, MA: MIT Press.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. New York: Oxford University Press.

Clark, A. (2006). Material Symbols. *Philosophical Psychology, 19*(3), pp. 291 - 307.

Clark, A. (2007). Curing cognitive hiccups: A defense of the extended mind. *The Journal of Philosophy, 104*(4), pp. 163-192.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

Clark, A. (2011). Finding the Mind. *Philosophical Studies, 152*(3), pp. 447-461.

Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences, 36*(3), pp. 181-204.

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis, 58*(1), pp. 7-19.

Clark, A., & Karmiloff-Smith, A. (1993). The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought. *Mind & Language, 8*(4), pp. 487-519.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*(3), pp. 401-431.

Cleeremans, A. (1993). *Mechanisms of implicit learning. Connectionist models of sequence processing*. Cambridge, MA: MIT Press.

Cleeremans, A. (1997). Principles for Implicit Learning. In D. Berry (Ed.), *How implicit is implicit learning?* Oxford: Oxford University Press, pp. 196-234.

Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective. In Q. Jing, M.R. Rosenzweig, G. d'Ydewalle, H. Zhang, H.-C. Chen & K. Zhang (Eds.), *Progress in Psychological Science around the world. Vol I. Neural, Cognitive and Developmental Issues* Hove: Psychology Press, pp. 401-418.

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R.M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness* Hove: Psychology Press, pp. 1–40.

Cleeremans, A., Timmermans, B., & Pasquali, A. (2007). Consciousness and metarepresentation: A computational sketch. *Neural Networks, 20*(9), pp. 1032-1039.

Clune, J., Mouret, J.-B., & Lipson, H. (2013). The evolutionary origins of modularity. *Proceedings of the Royal Society B: Biological Sciences, 280*(1755).

Coch, D., & Fischer, K.W. (1998). Discontinuity and variability in relational complexity: Cognitive and brain development. *Behavioral and Brain Sciences, 21*(06), pp. 834-835.

Cohen, G. (2000). Hierarchical models in cognition: Do they have psychological reality? *European Journal of Cognitive Psychology, 12*(1), pp. 1 - 36.

Cohen, N.J., Ryan, J., Hunt, C., et al. (1999). Hippocampal system and declarative (relational) memory: Summarizing the data from functional neuroimaging studies. *Hippocampus, 9*(1), pp. 83-98.

Collier, J. (1998). Information increase in biological systems: how does adaptation fit. *Evolutionary Systems*, pp. 129-140.

Colombo, M., & Seriès, P. (2012). Bayes in the Brain - On Bayesian Modelling in Neuroscience.

*British Journal for the Philosophy of Science, 63*(3), pp. 697-723.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences, 3*(3), pp. 115-120.

Cooper, R. (2002). Order and Disorder in Everyday Action: the Roles of Contention Scheduling and Supervisory Attention. *Neurocase, 8*(1-2), pp. 61-79.

Cooper, R., & Glasspool, D. (2001). Learning action affordances and action schemas. In R.M. French & J.P. Sougné (Eds.), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop* London: Springer, pp. 133-143.

Cooper, R., & Shallice, T. (1997). Modelling the selection of routine action: Exploring the criticality of parameter values. In Shafto, M.G. & Langley, P. (Eds.) *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, Palo Alto CA, pp. 131-136.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology, 17*(4), pp. 297-338.

Cooper, R.P. (2003). Mechanisms for the generation and regulation of sequential behaviour. *Philosophical Psychology, 16*(3), pp. 389-416.

Cooper, R.P., Schwartz, M.F., Yule, P., et al. (2005). The simulation of action disorganisation in complex activities of daily living. *Cognitive Neuropsychology, 22*(8), pp. 959-1004.

Cooper, R.P., & Shallice, T. (2006). Hierarchical Schemas and Goals in the Control of Sequential Behavior. *Psychological Review, 113*(4), pp. 887-916.

Cosmides, L., & Tooby, J. (1997). Evolutionary psychology: A primer. Retrieved on August 30, 2013 from http://www.psych.ucsb.edu/research/cep/primer.html

Courchesne, E., & Pierce, K. (2005). Why the frontal cortex in autism might be talking only to itself: local over-connectivity but long-distance disconnection. *Current Opinion in Neurobiology, 15*(2), pp. 225-230.

Craver, C.F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science, 68*(1), pp. 53-74.

Craver, C.F. (2002). Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory. *Philosophy-of-Science, 69*(3 Supplement), pp. S83-S97.

Craver, C.F. (2007). *Explaining the Brain. Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

Craver, C.F., & Darden, L. (2001). Discovering Mechanisms in Neurobiology: The Case of Spatial Memory. In P.K. Machamer (Ed.), *Theory and Method in the Neurosciences* Univ-of-Pitt-Pr: Pittsburgh 2001, pp. 112-137.

Creem, S.H. (2001). Defining the cortical visual systems: "what", "where", and "how". *Acta Psychologica, 107*(1-3), pp. 43-68.

Crescentini, C., Seyed-Allaei, S., De Pisapia, N., et al. (2011). Mechanisms of Rule Acquisition

and Rule Following in Inductive Reasoning. *Journal of Neuroscience, 31*(21), pp. 7763-7774.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars of the Neurosciences, 2*, pp. 263-275.

Crick, F., & Koch, C. (1995). Are we aware of neural activity in primary visual cortex? *Nature, 375*(6527), pp. 121-123.

Crisafi, A., & Gallagher, S. (2010). Hegel and the extended mind. *AI & Society, 25*(1), pp. 123-129.

Crowe, M.J. (1988). Ten Misconceptions about Mathematics and Its History. In W. Aspray & P. Kitcher (Eds.), *History and Philosophy of Modern Mathematics*: University of Minnesota Press.

Crumley, C.L. (1995). Heterarchy and the Analysis of Complex Societies. *Archeological Papers of the American Anthropological Association, 6*(1), pp. 1-5.

Crystal, J.D. (2013). Remembering the past and planning for the future in rats. *Behavioural Processes, 93*(0), pp. 39-49.

Csibra, G., & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica, 124*(1), pp. 60-78.

Cunningham, W.A., Zelazo, P.D., Packer, D.J., et al. (2007). The Iterative Reprocessing Model: A Multilevel Framework for Attitudes and Evaluation. *Social Cognition, 25*(5), pp. 736-760.

Damasio, A.R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*(1), pp. 25-62.

Daselaar, S.M., Porat, Y., Huijbers, W., et al. (2010). Modality-specific and modality-independent components of the human imagery system. *NeuroImage, 52*(2), pp. 677-685.

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy, 60*(23), pp. 685-700.

Davies, M. (2010). Double Dissociation: Understanding its Role in Cognitive Neuropsychology. *Mind & Language, 25*(5), pp. 500-540.

Dawkins, R., & Dawkins, M. (1976). Hierachical organization and postural facilitation: Rules for grooming in flies. *Animal Behaviour, 24*(4), pp. 739-755.

De Brigard, F., Addis, D.R., Ford, J.H., et al. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia, 51*(12), pp. 2401-2414.

De Brigard, F., & Giovanello, K.S. (2012). Influence of outcome valence in the subjective experience of episodic past, future, and counterfactual thinking. *Consciousness and Cognition, 21*(3), pp. 1085-1096.

De Cruz, H. (2008). An Extended Mind Perspective on Natural Number Representation. *Philosophical Psychology, 21*(4), pp. 475 - 490.

de Groot, A.D. (1946). *Het denken van den schaker: een experimenteel-psychologische studie.*

Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

de Oliveira, H., Cuervo-Lombard, C., Salame, P., et al. (2009). Autonoetic awareness associated with the projection of the self into the future: An investigation in schizophrenia. *Psychiatry Research, 169*(1), pp. 86-87.

de Vignemont, F., & Haggard, P. (2008). Action observation and execution: What is shared? *Social Neuroscience, 3*(3), pp. 421 - 433.

de Vries, S., & Mulder, T. (2007). Motor imagery and stroke rehabilitation: A critical discussion. *Journal of Rehabilitation Medicine, 39*(1), pp. 5-13.

Deacon, T. (2006). The Evolution of Language Systems in the Human Brain. In J. Kaas (Ed.), *Evolution of Nervous Systems. Volume 5 - The Evolution of Primate Nervous Systems* Oxford: Elsevier, pp. 81-106.

Deacon, T.W. (1997). *The symbolic species. The co-evolution of language and the brain*. New York, N.Y.: Norton.

Decety, J., Grezes, J., Costes, N., et al. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain, 120 ( Pt 10)*, pp. 1763-1777.

Dehaene, S. (2001). Precis of The Number Sense. *Mind & Language, 16*(1), pp. 16-36.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In S. Dehaene, J.R. Duhamel, M. Hauser & G. Rizzolatti (Eds.), *From monkey brain to human brain* Cambridge, MA: MIT Press, pp. 133-158.

Dehaene, S., Changeux, J.-P., Naccache, L., et al. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences, 10*(5), pp. 204-211.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition, 79*, pp. 1-37.

Dehaene, S., Spelke, E., Pinel, P., et al. (1999). Sources of Mathematical Thinking: Behavioral and Brain-Imaging Evidence. *Science, 284*(5416), pp. 970-974.

Demoss, D., & Devereux, D. (1988). Essence, Existence, and Nominal Definition in Aristotle's "Posterior Analytics" II 8-10. *Phronesis, 33*(2), pp. 133-154.

Dennett, D. (1978). *Brainstorms*: Bradford Books.

Dennett, D. (1987). Cognitive wheels: the frame problem of AI. In Z.W. Pylyshyn (Ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* Norwood, NJ: Ablex publishing, pp. 41-64.

Dennett, D. (2000). Making tools for thinking. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* Oxford: Oxford University Press, pp. 17-29.

Dennett, D. (2005). *Sweet Dreams. Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press.

Dennett, D.C. (1989). *The intentional stance*. Cambridge, MA: The MIT Press.

Dennett, D.C. (1993). *Consciousness Explained*. London: Penguin books.

Depew, D.J. (1991). Politics, Music, and Contemplation in Aristotle's Best State. In D. Keyt & F. Miller (Eds.), *A companion to Aristotle's Politics* Oxford: Blackwell.

Deutsch, R., & Strack, F. (2006). Duality Models in Social Psychology: From Dual Processes to Interacting Systems. *Psychological Inquiry, 17*(3), pp. 166-172.

Dewey, J. (1922). *Human Nature and Conduct: An Introduction to Social Psychology.* New York: Modern Library.

Di Nucci, E. (2011). Frankfurt versus Frankfurt: a new anti-causalist dawn. *Philosophical Explorations, 14*(1), pp. 117-131.

Di Paolo, E.A., Rohde, M., & De Jaegher, H. (2010). Horizons for the Enactive Mind: Values, Social Interaction, and Play. In J. Stewart, O. Gapenne & E.A. Di Paolo (Eds.), *Enaction: Towards a New Paradigm for Cognitive Science* Cambridge, MA: MIT Press, pp. 33-87.

Didierjean, A., & Gobet, F. (2008). Sherlock Holmes - an expert's view of expertise. *British Journal of Psychology, 99*, pp. 109-125.

Didierjean, A., & Marmèche, E. (2005). Anticipatory representation of visual basketball scenes by novice and expert players. *Visual Cognition, 12*(2), pp. 265-283.

Dieks, D. (2011). Quantum Mechanics, Chance and Modality. *Philosophica, 83*(1), pp. 117-137.

Dijksterhuis, A. (2007). *Het slimme onbewuste. Denken met gevoel*. Amsterdam: Bert Bakker.

Dijksterhuis, A., & Bargh, J.A. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology, 33*, pp. 1-40.

Dijksterhuis, A., Bos, M.W., Nordgren, L.F., et al. (2006). On Making the Right Choice: The Deliberation-Without-Attention Effect. *Science, 311*(5763), pp. 1005-1007.

Dijksterhuis, A., & Nordgren, L.F. (2006). A Theory of Unconscious Thought. *Perspectives on Psychological Science, 1*(2), pp. 95-109.

Dijksterhuis, E.J. (1969). *The mechanization of the world picture* (C. Dikshoorn, Trans.). Oxford: Oxford University Press.

Dominey, P.F., Lelekov, T., Ventre-Dominey, J., et al. (1998). Dissociable Processes for Learning the Surface Structure and Abstract Structure of Sensorimotor Sequences. *Journal of Cognitive Neuroscience, 10*(6), pp. 734-751.

Dominguez Duque, J.F., Lewis, E.D., Turner, R., et al. (2009). The brain in culture and culture in the brain: a review of core issues in neuroanthropology. *Progress in Brain Research 178*, pp. 43-64.

Donald, M. (1991). *Origins of the modern mind. Three stages in the evolution of culture and cognition*. Cambridge, MA: Harvard University Press.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition, 110*(3), pp. 412-431.

Dretske, F. (2003). Experience as Representation. *Philosophical Issues, 13*(1), pp. 67-82.

Dreyfus, H.L. (2002). Intelligence without representation – Merleau-Ponty's critique of mental representation. The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1*(4), pp. 367-383.

Dreyfus, H.L. (2009). How Representational Cognitivism Failed and is being replaced by Body/World Coupling. In L. K. (Ed.), *After Cognitivism* Dordrecht: Springer, pp. 39-73.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press.

Dreyfus, S.E. (2004). The Five-Stage Model of Adult Skill Acquisition. *Bulletin of Science Technology Society, 24*(3), pp. 177-181.

Dupré, J. (2001). In defence of classification. *Studies in History and Philosophy of Science Part C: Studies in Hist. and Philos. of Biological and Biomedical Sciences, 32*(2), pp. 203-219.

Durstewitz, D., Vittoz, N.M., Floresco, S.B., et al. (2010). Abrupt Transitions between Prefrontal Neural Ensemble States Accompany Behavioral Transitions during Rule Learning. *Neuron, 66*(3), pp. 438-448.

Durston, S., & Casey, B.J. (2006). What have we learned about cognitive development from neuroimaging? *Neuropsychologia, 44*(11), pp. 2149-2157.

Dusek, J.A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *PNAS, 94*(13), pp. 7109-7114.

Egidi, G., & Caramazza, A. (2013). Cortical systems for local and global integration in discourse comprehension. *NeuroImage, 71*(0), pp. 59-74.

Elk, M., Viswanathan, S., Schie, H.T., et al. (2012). Pouring or chilling a bottle of wine: an fMRI study on the prospective planning of object-directed actions. *Experimental Brain Research, 218*(2), pp. 189-200.

Elman, J.L., Bates, E., Johnson, M., et al. (1997). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: The MIT Press.

Emery, N.J., & Clayton, N.S. (2009). Tool use and physical cognition in birds and mammals. *Current Opinion in Neurobiology, 19*(1), pp. 27-33.

Ericsson, K.A., Krampe, R.T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, pp. 363-363.

Ericsson, K.A., & Roring, R.W. (2007). Memory as A Fully Integrated Aspect of Skilled and Expert Performance. In S.B. Aaron & H.R. Brian (Eds.), *Psychology of Learning and Motivation* Vol. Volume 48: Academic Press, pp. 351-380.

Ericsson, K.A., Roring, R.W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: an account based on the expert performance framework. *High Ability Studies, 18*(1), pp. 3-56.

Evans, J.S.B.T. (2003). In two minds: dual-process accounts of reasoning *Trends in Cognitive*

*Sciences 7*(10), pp. 454-459.

Evans, J.S.B.T. (2006). Dual system theories of cognition: Some issues. In R. Sun (Ed.), *Proceedings of the 28th annual meeting of the cognitive science society* Mahwah, NJ: Erlbaum, pp. 202–207.

Evans, J.S.B.T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology, 59*(1), pp. 255-278.

Evans, J.S.B.T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review, 31*(2-3), pp. 86-102.

Fadiga, L., Craighero, L., & D'Ausilio, A. (2009). Broca's Area in Language, Action, and Music. *Annals of the New York Academy of Sciences, 1169*(1), pp. 448-458.

Farne, A., Iriki, A., & Ladavas, E. (2005). Shaping multisensory action-space with tools: evidence from patients with cross-modal extinction. *Neuropsychologia, 43*(2), pp. 238-248.

Feist, G.J. (2013). The nature and nurture of expertise: a fourth dimension. *Phenomenology and the Cognitive Sciences, 13*(2), pp. 275-288.

Feldman Barrett, L., Tugade, M.M., & Engle, R.W. (2004). Individual Differences in Working Memory Capacity and Dual-Process Theories of the Mind. *Psychological Bulletin;Psychological Bulletin, 130*(4), pp. 553-573.

Felix, M.A., & Wagner, A. (2006). Robustness and evolution: concepts, insights and challenges from a developmental model system. *Heredity, 100*(2), pp. 132-140.

Felleman, D.J., & Van Essen, D.C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb. Cortex, 1*(1), pp. 1-a-47.

Ferrari, V., Didierjean, A., & Marmèche, E. (2008). Effect of expertise acquisition on strategic perception: The example of chess. *The Quarterly Journal of Experimental Psychology, 61*(8), pp. 1265-1280.

Fertonani, A., Pirulli, C., & Miniussi, C. (2011). Random Noise Stimulation Improves Neuroplasticity in Perceptual Learning. *The Journal of Neuroscience, 31*(43), pp. 15416-15423.

Fessler, D.M.T., & Navarrete, C.D. (2003). Meat Is Good to Taboo: Dietary Proscriptions as a Product of the Interaction of Psychological Mechanisms and Social Processes. *Journal of Cognition and Culture, 3*(1), pp. 1-40.

Fingelkurts, A., & Fingelkurts, A. (2006). Timing in cognition and EEG brain dynamics: discreteness versus continuity. *Cognitive Processing, 7*(3), pp. 135-162.

Fischer, J. (2004). Responsibility and Manipulation. *The Journal of Ethics, 8*(2), pp. 145-177.

Fischer, J.M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility.* Cambridge: Cambridge University Press.

Fitch, W.T., & Hauser, M.D. (2004). Computational Constraints on Syntactic Processing in a Nonhuman Primate. *Science, 303*(5656), pp. 377-380.

Fletcher, P.C., Shallice, T., & Dolan, R.J. (2000). Sculpting the Response Space. An Account of Left Prefrontal Activation at Encoding. *Neuroimage, 12*(4), pp. 404-417.

Fodor, J.A. (1983). *The modularity of mind. An essay on faculty psychology.* Cambridge, MA: MIT Press.

Forbes, C.E., & Grafman, J. (2010). The Role of the Human Prefrontal Cortex in Social Cognition and Moral Judgment. *Annual Review of Neuroscience, 33*, pp. 299-324.

Forman, D. (2008). Autonomy as Second Nature: On McDowell's Aristotelian Naturalism. *Inquiry, 51*(6), pp. 563-580.

Fourkas, A.D., Bonavolontà, V., Avenanti, A., et al. (2008). Kinesthetic Imagery and Tool-Specific Modulation of Corticospinal Representations in Expert Tennis Players. *Cerebral Cortex, 18*(10), pp. 2382-2390.

Fox, M.D., & Raichle, M.E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat Rev Neurosci, 8*(9), pp. 700-711.

Fox, M.D., Snyder, A.Z., Vincent, J.L., et al. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), pp. 9673-9678.

Franck, R. (2004). Should neuroconstructivism guide developmental research? *Trends in Cognitive Sciences, 8*(3), pp. 100-101.

Frankfurt, H. (2002a). Reply to Eleonore Stump. In S. Buss & L. Overton (Eds.), *Contours of agency: essays on themes from Harry Frankfurt* Cambridge, MA: MIT Press, pp. 61-63.

Frankfurt, H. (2002b). Reply to Susan Wolf. In S. Buss & L. Overton (Eds.), *Contours of agency: essays on themes from Harry Frankfurt* Cambridge, MA: MIT Press, pp. 245-252.

Frankfurt, H.G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy, LXVIII*(1), pp. 5-20.

Frankfurt, H.G. (1978). The Problem of Action. *American philosophical quarterly, 15*(2), pp. 157-162.

Frankfurt, H.G. (1988). *The importance of what we care about.* Cambridge: Cambridge University Press.

Frankfurt, H.G. (1999). *Necessity, volition, and love.* Cambridge: Cambridge University Press.

Frankfurt, H.G. (2002). Reply to Michael E. Bratman. In S. Buss & L. Overton (Eds.), *Contours of agency: essays on themes from Harry Frankfurt* Cambridge, MA: MIT Press, pp. 86-90.

Frankish, K. (2010). Dual-Process and Dual-System Theories of Reasoning. *Philosophy Compass, 5*(10), pp. 914-926.

Frankish, K., & Evans, J.S.B.T. (2009). The duality of mind: An historical perspective. In J.S.B.T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* New York: Oxford University Press, pp. 1-29.

Freeman, W.J. (2000). Mesoscopic neurodynamics: From neuron to brain. *Journal of*

*Physiology-Paris, 94*(5-6), pp. 303-322.

Freeth, T., Bitsakis, Y., Moussas, X., et al. (2006). Decoding the ancient Greek astronomical calculator known as the Antikythera Mechanism. *Nature, 444*(7119), pp. 587-591.

Frijda, N.H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1456), pp. 815-836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci, 11*(2), pp. 127-138.

Frith, C.D. (2000). The role of dorsolateral prefrontal cortex in the selection of action as revealed by functional imaging. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* Cambridge, MA: MIT Press pp. 547-565.

Frith, C.D., Friston, K., Liddle, P.F., et al. (1991). Willed action and the prefrontal cortex in man: a study with PET. *Proceedings Of The Royal Society Of London. Series B. Biological Sciences, 244*(1311), pp. 241.

Frye, D., & Zelazo, P.D. (1998). Complexity: From formal analysis to final action. *Behavioral and Brain Sciences, 21*(06), pp. 836-837.

Fusaroli, R., Bahrami, B., Olsen, K., et al. (2012). Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science, 23*(8), pp. 931-939.

Fuster, J. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe* Philadelphia: Lippincott Williams & Wilkins.

Fuster, J.M. (2000). The prefrontal cortex of the primate: A synopsis. *Psychobiology, 28*(2), pp. 125-131.

Fuster, J.M. (2001). The Prefrontal Cortex--An Update: Time Is of the Essence. *Neuron, 30*(2), pp. 319-333.

Galinsky, A.D., & Moskowitz, G.B. (2000). Counterfactuals as Behavioral Primes: Priming the Simulation Heuristic and Consideration of Alternatives. *Journal of Experimental Social Psychology, 36*(4), pp. 384-409.

Gallagher, S. (2005). *How the body shapes the mind*. Oxford: Clarendon Press.

Gallagher, S. (2012). Multiple aspects in the sense of agency. *New Ideas in Psychology, 30*(1), pp. 15-31.

Gallagher, S., & Hutto, D. (2008). Understanding others through primary interaction and narrative practice. In J. Zlatev, T. Racine, C. Sinha & E. Itkonen (Eds.), *The shared mind: Perspectives on intersubjectivity* Amsterdam: John Benjamins, pp. 17–38.

Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. *Social Neuroscience, 3*(3), pp. 317 - 333.

Gallese, V., Fadiga, L., Fogassi, L., et al. (1996). Action recognition in the premotor cortex. *Brain, 119*(2), pp. 593-609.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), pp. 493-501.

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22*(3), pp. 455 - 479.

Garbin, G., Sanjuan, A., Forn, C., et al. (2010). Bridging language and attention: Brain basis of the impact of bilingualism on cognitive control. *NeuroImage, 53*(4), pp. 1272-1278.

Gärdenfors, P. (1996). Mental representation, conceptual spaces and metaphors. *Synthese, 106*(1), pp. 21-47.

Gärdenfors, P. (2004a). Conceptual Spaces as a Framework for Knowledge Representation. *Mind and Matter, 2*, pp. 9-27.

Gärdenfors, P. (2004b). *Conceptual Spaces: The Geometry of Thought.* Cambridge, MA: MIT Press.

Gärdenfors, P., & Williams, M.A. (2003). Building rich and grounded robot world models from sensors and knowledge resources: A conceptual spaces approach *Proceedings of AMIRE 2003* Brisbane: Queensland University of Technology, pp. 123-132.

Genty, E., Breuer, T., Hobaiter, C., et al. (2009). Gestural communication of the gorilla (Gorilla gorilla): repertoire, intentionality and possible origins. *Animal Cognition, 12*(3), pp. 527-546.

Georgieff, N., & Jeannerod, M. (1998). Beyond Consciousness of External Reality: A "Who" System for Consciousness of Action and Self-Consciousness. *Consciousness and Cognition, 7*(3), pp. 465-477.

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415*(6873), pp. 755.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences, 7*(7), pp. 287-292.

Gerlach, K.D., Dornblaser, D.W., & Schacter, D.L. (2013). Adaptive constructive processes and memory accuracy: Consequences of counterfactual simulations in young and older adults. *Memory*, pp. 1-18.

Gerlach, K.D., Spreng, R.N., Gilmore, A.W., et al. (2011). Solving future problems: Default network and executive activity associated with goal-directed mental simulations. *NeuroImage, 55*(4), pp. 1816-1824.

Gerrig, R.J. (1993). *Experiencing narrative worlds: On the psychological activities of reading* New Haven CT: Yale University Press.

Ghio, M., & Tettamanti, M. (2010). Semantic domain-specific functional integration for action-related vs. abstract concepts. *Brain and Language, 112*(3), pp. 223-232.

Gilbert, D.T. (1999). What the Mind's not. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* New York: Guilford Press.

Gilman, D. (1994). Simplicity, Cognition and Adaptation: Some Remarks on Marr's Theory of Vision. *Proceedings of the biennial meeting of the Philosophy of Science Association*, pp. 454-464.

Giora, R., & Shen, Y. (1994). Degrees of narrativity and strategies of semantic reduction. *Poetics, 22*(6), pp. 447-458.

Glenberg, A.M. (2011). Introduction to the Mirror Neuron Forum. *Perspectives on Psychological Science, 6*(4), pp. 363-368.

Glenberg, A.M., Sato, M., Cattaneo, L., et al. (2008). Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology, 61*(6), pp. 905 - 919.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*.

Glennan, S. (2008). Mechanisms. In S. Psillos & M. Curd (Eds.), *The Routledge Companion to the Philosophy of Science* London: Routledge, pp. 376-384.

Glennerster, A. (2002). Computational theories of vision. *Current Biology, 12*(20), pp. R682-R685.

Glennerster, A. (2007). Marr's vision: twenty-five years on. C*urrent Biology, 17*(11), pp. R397.

Gobet, F. (1998). Expert memory: a comparison of four theories. *Cognition, 66*(2), pp. 115-152.

Gobet, F., & Chassy, P. (2009). Expertise and Intuition: A Tale of Three Theories. *Minds and Machines, 19*(2), pp. 151-180.

Gobet, F., Lane, P.C.R., Croker, S., et al. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), pp. 236-243.

Gobet, F., & Simon, H.A. (1996). Templates in Chess Memory: A Mechanism for Recalling Several Boards. *Cognitive Psychology, 31*(1), pp. 1-40.

Gobet, F., & Simon, H.A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24*(4), pp. 651-682.

Godbout, L., Cloutier, P., Bouchard, C., et al. (2004). Script Generation Following Frontal and Parietal Lesions. *Journal of Clinical and Experimental Neuropsychology, 26*(7), pp. 857-873.

Goldman, A. (2006). *Simulating Minds. The Philosophy, Psychology, and Neuroscience of Mindreading.* New York: Oxford University Press.

Goldstone, R.L., & Barsalou, L.W. (1998). Reuniting perception and conception. *65*, pp. 231-262.

Goldstone, R.L., Landy, D.H., & Son, J.Y. (2010). The Education of Perception. *Topics in Cognitive Science, 2*(2), pp. 265-284.

Gollwitzer, P.M. (1993). Goal Achievement: The Role of Intentions. *European Review of Social Psychology, 4*(1), pp. 141-185.

Gollwitzer, P.M., & Brandstatter, V. (1997). Implementation Intentions and Effective Goal Pursuit. *Journal of Personality & Social Psychology, 73*(1), pp. 186-199.

Gollwitzer, P.M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. In M.P. Zanna (Ed.), *Advances in experimental social psychology* Vol. 38, San Diego, CA: Academic Press, pp. 69-119.

Goodale, M.A., & Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), pp. 20-25.

Gotthelf, A. (1987). First Principles in Aristotle's Parts of Animals. In A. Gotthelf & J. Lennox (Eds.), *Philosophical Issues in Aristotle Biology* Cambridge: Cambridge University Press.

Gotthelf, A. (1997). The elephant's nose: further reflections on the axiomatic structure of biological explanation in Aristotle. In W. Kullmann & S. Föllinger (Eds.), *Aristotelische Biologie : Intentionen, Methoden, Ergebnisse* Stuttgart: Steiner Verlag, pp. 85-95.

Grafman, J. (1995). Similarities and Distinctions among Current Models of Prefrontal Cortical Functions. *Annals of the New York Academy of Sciences, 769*(1), pp. 337-368.

Grafman, J. (2003). The structured event complex and the human prefrontal cortex. *Cognitive Processes and Economic Behaviour*, pp. 209.

Grafman, J. (2006). Human prefrontal cortex: processes and representations. In J. Risberg & J. Grafman (Eds.), *The Frontal Lobes. Development, Function, and Pathology* Cambridge: Cambridge University Press, pp. 69-91.

Grafman, J., & Krueger, F. (2008). The prefrontal cortex stores structured event complexes that are the representational basis for cognitively derived actions. In E. Morsella, J.A. Bargh & P.M. Gollwitzer (Eds.), *Oxford Handbook of Human Action: Mechanisms of Human Action* New York: Oxford University Press, pp. 197-213.

Grafman, J., Sirigu, A., Spector, L., et al. (1993). Damage to the prefrontal cortex leads to decomposition of structured event complexes. *The Journal of Head Trauma Rehabilitation, 8*(1), pp. 73-87.

Grafton, S.T., & de C. Hamilton, A.F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science, 26*(4), pp. 590-616.

Graybiel, A.M. (1998). The Basal Ganglia and Chunking of Action Repertoires. *Neurobiology of Learning and Memory, 70*(1-2), pp. 119-136.

Graybiel, A.M. (2008). Habits, Rituals, and the Evaluative Brain. *Annual Review of Neuroscience, 31*(1), pp. 359-387.

Graybiel, A.M., & Rauch, S.L. (2000). Toward a Neurobiology of Obsessive-Compulsive Disorder. *Neuron, 28*(2), pp. 343-347.

Greenfield, P.M. (1991). From hand to mouth. *Behavioral and Brain Sciences, 14*(04), pp. 577-595.

Gregor, B. (2005). Selfhood and the three R's: Reference, Repetition, and Refiguration. *International Journal for Philosophy of Religion, 58*(2), pp. 63-94.

Grezes, J., & Decety, J. (2001). Functional anatomy of execution, mental simulation,

observation, and verb generation of actions: A meta-analysis. *Human Brain Mapping, 12*(1), pp. 1-19.

Grèzes, J., & Decety, J. (2002). Does visual perception of object afford action? Evidence from a neuroimaging study. *Neuropsychologia, 40*(2), pp. 212-222.

Grice, P. (1974). Method in Philosophical Psychology (From the Banal to the Bizarre). *Proceedings and Addresses of the American Philosophical Association, 48*, pp. 23-53.

Griffiths, P.E. (2007). Evo-Devo meets the mind: Towards a developmental evolutionary psychology. In R. Sansom & R.N. Brandon (Eds.), *Integrating evolution and development: from theory to practice* Cambridge, MA: MIT press, pp. 195-226.

Grimm, S., Ernst, J., Boesiger, P., et al. (2009). Increased Self-Focus in Major Depressive Disorder Is Related to Neural Abnormalities in Subcortical-Cortical Midline Structures. *Human Brain Mapping, 30*(8), pp. 2617-2627.

Gross, C.G. (2002). Genealogy of the "Grandmother Cell". *Neuroscientist, 8*(5), pp. 512-518.

Gruber, M.J., & Otten, L.J. (2010). Voluntary Control over Prestimulus Activity Related to Encoding. *J. Neurosci., 30*(29), pp. 9793-9800.

Guida, A., Gobet, F., Tardieu, H., et al. (2012). How chunks, long-term working memory and templates offer a cognitive explanation for neuroimaging data on expertise acquisition: A two-stage framework. *Brain and Cognition, 79*(3), pp. 221-244.

Gyurak, A., Gross, J.J., & Etkin, A. (2011). Explicit and implicit emotion regulation: A dual-process framework. *Cognition & Emotion, 25*(3), pp. 400-412.

Hacker, P.M.S., & Bennett, M.R. (2011). Isms are prisms: a reply to Keestra and Cowley. *Language Sciences, 33*(3), pp. 459-463.

Hacking, I. (1991). A Tradition of Natural Kinds. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 61*(1/2), pp. 109-126.

Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack & A.J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* Oxford: Clarendon Press, pp. 351-394.

Hagoort, P., & Levelt, W.J.M. (2009). The Speaking Brain. *Science, 326*(5951), pp. 372-373.

Halford, G.S., Wilson, W.H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*(06), pp. 803-831.

Halford, G.S., Wilson, W.H., & Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. *Trends in Cognitive Sciences, 14*(11), pp. 497-505.

Halliwell, S. (1987). *The Poetics of Aristotle: translation and commentary.* London: Duckworth.

Hameroff, S., & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation, 40*(3-4), pp. 453-480.

Hamzei, F., Rijntjes, M., Dettmers, C., et al. (2003). The human action recognition system and its relationship to Broca's area: an fMRI study. *NeuroImage, 19*(3), pp. 637-644.

Han, S., Gu, X., Mao, L., et al. (2010). Neural substrates of self-referential processing in Chinese Buddhists. *Social Cognitive and Affective Neuroscience, 5*, pp. 332-339.

Han, S.H., & Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: a transcultural neuroimaging approach. *Nature Reviews Neuroscience, 9*(8), pp. 646-654.

Hard, B.M., Lozano, S.C., & Tversky, B. (2006). Hierarchical Encoding of Behavior: Translating Perception Into Action. *Journal of Experimental Psychology: General, 135*(4), pp. 588-608.

Hardcastle, V.G. (2000). How to understand the N in NCC. In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions. .* Cambridge, MA: MIT Press, pp. 259-264.

Hardt, O., Einarsson, E., & Nader, K. (2010). A Bridge Over Troubled Water: Reconsolidation as a Link Between Cognitive and Neuroscientific Memory Research Traditions. *Annual Review of Psychology, 61*(1), pp. 141-167.

Harmelech, T., Preminger, S., Wertman, E., et al. (2013). The Day-After Effect: Long Term, Hebbian-Like Restructuring of Resting-State fMRI Patterns Induced by a Single Epoch of Cortical Activation. *The Journal of Neuroscience, 33*(22), pp. 9488-9497.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), pp. 335-346.

Hartmann, N. (1957 [1923]). Aristoteles und Hegel *Kleinere Schriften II* Berlin: De Gruyter.

Haselager, P., de Groot, A., & van Rappard, H. (2003). Representationalism vs. anti-representationalism: a debate for the sake of appearance. *Philosophical Psychology, 16*, pp. 5-24.

Hassabis, D., & Maguire, E.A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences, 11*(7), pp. 299-306.

Hassabis, D., Spreng, R.N., Rusu, A.A., et al. (2013). Imagine All the People: How the Brain Creates and Uses Personality Models to Predict Behavior. *Cerebral Cortex*.

Hassin, R.R., Bargh, J.A., & Zimerman, S. (2009). Automatic and flexible: The case of non-conscious goal pursuit. *Social cognition, 27*(1), pp. 20-36.

Hatfield, G. (1991). Representation in Perception and Cognition: Connectionist Affordances. In W.M. Ramsey, S.P. Stich & D.E. Rumelhart (Eds.), *Philosophy and Connectionist Theory* Hillsdale, NJ: Lawrence Erlbaum, pp. 129.

Hauser, M.D., Chomsky, N., & Fitch, W.T. (2002). The Faculty of Language: What Is It, Who has it, and How Did It Evolve? *Science, 298*, pp. 1569-1579.

Hedström, P. (2008). Studying mechanisms to strengthen causal inferences in quantitative research. In J.M. Box-Steffensmeier, H.E. Brady & D. Collier (Eds.), *The Oxford Handbook of Political Methodology* Oxford: Oxford University Press, pp. 319-335.

Hedström, P., & Swedberg, R. (1996). Social Mechanisms. *Acta Sociologica, 39*(3), pp. 281-308.

Hedström, P., & Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology, 36*(1), pp. 49-67.

Hegel, G.W.F. (1988). *Phänomenologie des Geistes*. Hamburg: Meiner Verlag.

Henrich, J., Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), pp. 61-83.

Herman, D. (2003). Stories as a tool for thinking. In D. Herman (Ed.), *Narrative theory and the cognitive sciences*. Stanford, CA: CSLI Publications, pp. 163-192.

Herman, D. (2007). Introduction. In D. Herman (Ed.), *The Cambridge companion to narrative*. Cambridge: Cambridge University Press.

Herman, D. (2009). Storied Minds: Narrative Scaffolding for Folk Psychology. *Journal of Consciousness Studies, 16*(6-8), pp. 40-68.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6*(6), pp. 242-247.

Hesslow, G. (2012). The current status of the simulation theory of cognition. *Brain Research, 1428*, pp. 71-79.

Heyder, K., Suchan, B., & Daum, I. (2004). Cortico-subcortical contributions to executive control. *Acta Psychologica, 115*(2‚Äì3), pp. 271-289.

Hodgkinson, G.P., Langan-Fox, J., & Sadler-Smith, E. (2008). Intuition: A fundamental bridging construct in the behavioural sciences. *British Journal of Psychology, 99*, pp. 1-27.

Hofmann, W., Friese, M., & Strack, F. (2009). Impulse and Self-Control From a Dual-Systems Perspective. *Perspectives on Psychological Science, 4*(2), pp. 162-176.

Hofmann, W., Friese, M., & Wiers, R.W. (2008). Impulsive versus reflective influences on health behavior: a theoretical framework and empirical review. *Health Psychology Review, 2*(2), pp. 111-137.

Hohwy, J. (2007). Functional integration and the mind. *Synthese, 159*(3), pp. 315-328.

Hohwy, J., & Frith, C. (2004). Can Neuroscience Explain Consciousness? *Journal of Consciousness Studies, 11*, pp. 180-198.

Hollis, S., & Low, J. (2005). Karmiloff-Smith's RRM distinction between adjunctions and redescriptions: It's about time (and children's drawings). *British Journal of Developmental Psychology, 23*, pp. 623-644.

Holmes, N.P., Calvert, G.A., & Spence, C. (2004). Extending or projecting peripersonal space with tools? Multisensory interactions highlight only the distal and proximal ends of tools. *Neuroscience Letters, 372*(1–2), pp. 62-67.

Holton, R. (1999). Intention and Weakness of Will. *The Journal of Philosophy, 96*(5), pp. 241-262.

Holyoak, K.J., Koh, K., & Nisbett, R.E. (1989). A Theory of Conditioning: Inductive Learning

Within Rule-Based Default Hierarchies. *Psychological Review, 96*(2), pp. 315-340.

Hommel, B., Musseler, J., Aschersleben, G., et al. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral-and-Brain-Sciences, 24*(5), pp. 849-937.

Honing, H.J. (2009). *Iedereen is muzikaal. Wat we weten over het luisteren naar muziek.* Amsterdam: Nieuw Amsterdam

Huey, E., Zahn, R., Krueger, F., et al. (2008). A psychological and neuroanatomical model of obsessive-compulsive disorder. *The Journal of Neuropsychiatry and Clinical Neurosciences, 20*(4), pp. 390-408.

Hutchins, E. (2010). Cognitive Ecology. *Topics in Cognitive Science, 2*(4), pp. 705-715.

Hutchins, E., & Johnson, C.M. (2009). Modeling the Emergence of Language as an Embodied Collective Cognitive Activity. *Topics in Cognitive Science, 1*(3), pp. 523-546.

Hutchins, S., & Peretz, I. (2011). Perception and action in singing. In A.M. Green, C.E. Chapman, J.F. Kalaska & F. Lepore (Eds.), *Progress in Brain Research* Vol. Volume 191: Elsevier, pp. 103-118.

Hutto, D.D. (2007). *Folk psychological narratives: The sociocultural basis of understanding reasons.* Cambridge, MA: The MIT Press.

Hyman, S. (2011). Diagnosing the DSM. Diagnostic Classification Needs Fundamental Reform. *Cerebrum* Retrieved April, 29, 2011, from http://dana.org/news/cerebrum/detail.aspx?id=32066

Hyman, S.E. (2007). Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience, 8*(9), pp. 725-732.

Iacoboni, M. (2009). Imitation, Empathy, and Mirror Neurons. *Annual Review of Psychology, 60*(1), pp. 653-670.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., et al. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3*(3), pp. 0529-0535.

Imamizu, H., Kuroda, T., Miyauchi, S., et al. (2003). Modular organization of internal models of tools in the human cerebellum. *Proceedings of the National Academy of Sciences, 100*(9), pp. 5461-5466.

Ingvar, D.H. (1985). "Memory of the future": an essay on the temporal organization of conscious awareness. *Human neurobiology, 4*(3), pp. 127-136.

Iriki, A. (2006). The neural origins and implications of imitation, mirror neurons and tool use. *Current Opinion in Neurobiology, 16*(6), pp. 660.

Iriki, A., Tanaka, M., & Iwamura, Y. (1996). Coding of modified body schema during tool use by macaque postcentral neurones. *NeuroReport, 7*(14), pp. 2325-2330.

Isaac, A. (2009). Prospects for Naturalizing Color. *Philosophy Of Science, 76*(5), pp. 902-914.

Jacobson, A.J. (2003). Mental representations: what philosophy leaves out and neuroscience

puts in. *Philosophical Psychology, 16*(2), pp. 189-203.

Janelle, C.M., & Hillman, C.H. (2003). Expert Performance in sport. In J.L. Starkes & K.A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sport expertise*, pp. 19-47.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), pp. 187-245.

Jeannerod, M. (1997). *The cognitive neuroscience of action.* Oxford: Blackwell Publishers.

Jeannerod, M. (1999). To act or not to act: perspectives on the representation of actions. The 25th Bartlett Lecture. *Q J Exp Psychol A, 52(1):1-29.*

Jeannerod, M. (2001). Neural Simulation of Action: A Unifying Mechanism for Motor Cognition. *NeuroImage, 14*(1), pp. S103.

Jeannerod, M. (2006). *Motor Cognition. What actions tell the self.* New York: Oxford University Press.

Jeannerod, M. (2008). Language, perception and action. How words are grounded in the brain. *European Review, 16*(4), pp. 389-398.

Jeannerod, M., Arbib, M.A., Rizzolatti, G., et al. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences, 18*(7), pp. 314-320.

Jeannerod, M., & Frak, V. (1999). Mental imaging of motor activity in humans. *Current Opinion in Neurobiology, 9*(6), pp. 735.

Jeannerod, M., & Pacherie, E. (2004). Agency, Simulation and Self-identification. *Mind & language, 19*(2), pp. 113-146.

Jessup, R.K. (2009). Transfer of high domain knowledge to a similar domain. *The American journal of psychology, 122*(1), pp. 63-73.

Jirak, D., Menz, M.M., Buccino, G., et al. (2010). Grasping language - A short story on embodiment. *Consciousness and Cognition, 19*(3), pp. 711-720.

Johnson, J.G., & Raab, M. (2003). Take The First: Option-generation and resulting choices. *Organizational Behavior and Human Decision Processes, 91*(2), pp. 215-229.

Johnson, M.H., & Munakata, Y. (2005). Processes of change in brain and cognitive development. *Trends in Cognitive Sciences, 9*(3), pp. 152-158.

Johnson-Frey, S.H. (2003). What's So Special about Human Tool Use? *Neuron, 39*(2), pp. 201-204.

Johnson-Frey, S.H. (2004). The neural bases of complex tool use in humans. *Trends in Cognitive Sciences, 8*(2), pp. 71-78.

Juncos-Rabadán, O., Pereiro, A.X., & Rodríguez, M.S. (2005). Narrative speech in aging: Quantity, information content, and cohesion. *Brain and Language, 95*(3), pp. 423-434.

Justus, T.C., & Bharucha, J.J. (2001). Modularity in musical processing: The automaticity of harmonic priming. *Journal of Experimental Psychology, 27*(4), pp. 1000-1011.

Kandel, E.R. (2009). The Biology of Memory: A Forty-Year Perspective. *The Journal of Neuroscience, 29*(41), pp. 12748-12756.

Kantak, S.S., Sullivan, K.J., Fisher, B.E., et al. (2010). Neural substrates of motor memory consolidation depend on practice structure. *Nat Neurosci, 13,* pp. 923-935.

Kaplan, D.M., & Bechtel, W. (2011). Dynamical Models: An Alternative or Complement to Mechanistic Explanations? *Topics in Cognitive Science, 3*(2), pp. 438-444.

Karmiloff-Smith, A. (1990). Constraints on representational change: Evidence from children's drawing. *Cognition, 34*(1), pp. 57-83.

Karmiloff-Smith, A. (1992). *Beyond modularity. A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1994). Precis of: Beyond modularity: a developmental perspective on cognitive science. *Behavioral and Brain Sciences, 17*(4), pp. 693-745.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences, 2*(10), pp. 389-398.

Karmiloff-Smith, A. (2006). Ontogeny, Genetics, and Evolution: A Perspective from Developmental Cognitive Neuroscience. *Biological Theory, 1*(1), pp. 44-51.

Karmiloff-Smith, A. (2006). The tortuous route from genes to behavior: a neuroconstructivist approach. *Cognitive, affective, & behavioral neuroscience, 6*(1), pp. 9-17.

Karmiloff-Smith, A. (2009). Nativism Versus Neuroconstructivism: Rethinking the Study of Developmental Disorders. *Developmental Psychology, 45*(1), pp. 56-63.

Karmiloff-Smith, A. (2011). Static Snapshots versus Dynamic Approaches to Genes, Brain, Cognition, and Behavior in Neurodevelopmental Disabilities. In D.J. Fidler (Ed.), *Early Development in Neurogenetic Disorders* London: Elsevier, pp. 1-15.

Karmiloff-Smith, A., Scerif, G., & Ansari, D. (2003). Double Dissociations in Developmental Disorders? Theoretically Misconceived, Empirically Dubious. *Cortex, 39*(1), pp. 161-163.

Kay, P., Berlin, B., & Merrifield, W. (1991). Biocultural Implications of Systems of Color Naming. *Journal of Linguistic Anthropology, 1*(1), pp. 12-25.

Keele, S.W., Jennings, P., Jones, S., et al. (1995). On the Modularity of Sequence Representation. *Journal of Motor Behavior, 27*(1), pp. 17-30.

Keestra, M. (1991). Waar stil te staan? Aristoteles en de vraag naar principes van kennis. *Stoicheia : tijdschrift voor historische wijsbegeerte.* Vol. 6, pp. 3-24.

Keestra, M. (2000). Aristoteles. In M. Keestra (Ed.), *Tien westerse filosofen* Amsterdam: Nieuwezijds, pp. 47-67.

Keestra, M. (2008). The Diverging Force of Imitation: Integrating Cognitive Science and Hermeneutics. *Review of General Psychology, 12*(2), pp. 127-136.

Keestra, M. (2011). Understanding human action: integrating meanings, mechanisms, causes, and contexts. In A. Repko, W.H. Newell & R. Szostak (Eds.), *Interdisciplinary research:*

*Case studies of interdisciplinary understandings of complex problems.* Thousand Oaks, CA: SAGE publications, pp. 225-258.

Keestra, M. (2012). Bounded mirroring: joint action and group membership in political theory and cognitive neuroscience. In F. Vandervalk (Ed.), *Thinking about the Body Politic: Essays on Neuroscience and Political Theory.* London: Routledge, pp. 222-248.

Keestra, M., & Cowley, S.J. (2009). Foundationalism and neuroscience; silence and language. *Language Sciences, 31*(4), pp. 531-552.

Keestra, M., & Cowley, S.J. (2011). Concepts – not just yardsticks, but also heuristics: rebutting Hacker and Bennett. *Language Sciences, 33*(3), pp. 464-472.

Keijzer, F. (2002). Representation in dynamical and embodied cognition. *Cognitive Systems Research, 3*(3), pp. 275-288.

Kelso, J.A.S. (2009). Synergies: Atoms of Brain and Behavior. In D. Sternad (Ed.), *Progress in Motor Control.* In Vol. 629: Springer US, pp. 83-91.

Kenny, A. (1979). *Aristotle's Theory Of The Will.* London: Durckworth.

Kessler, M. (1976). *Aristoteles' Lehre von der Einheit der Definition.* München: J. Berchmans.

Keysers, C., & Gazzola, V. (2009). Unifying Social Cognition. In J.A. Pineda (Ed.), *Mirror Neuron Systems* New York, NY: Humana Press, pp. 1-35.

Keysers, C., & Perrett, D.I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences, 8*(11), pp. 501-507.

Kim, C., Johnson, N.F., Cilles, S.E., et al. (2011). Common and Distinct Mechanisms of Cognitive Flexibility in Prefrontal Cortex. *Journal of Neuroscience, 31*(13), pp. 4771-4779.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research, 52*(1), pp. 1-26.

Kim, J. (2000). *Mind in a physical world: An essay on the mind-body problem and mental causation.* Cambridge, MA: The MIT press.

Kinsbourne, M., & Jordan, J.S. (2009). Embodied Anticipation: A Neurodevelopmental Interpretation. *Discourse Processes, 46*(2), pp. 103 - 126.

Kirchhoff, M. (2012). Extended cognition and fixed properties: steps to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences, 11*(2), pp. 287-308.

Kitcher, P. (1988). Marr's computational theory of vision. *Philosophy of science : official journal of the Philosophy of Science Association, 55*(1), pp. 1-24.

Klein, G.A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid Decision Making on the Fire Ground. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 30*(6), pp. 576-580.

Klein, K. (2003). Narrative construction, cognitive processing, and health. In D. Herman (Ed.), *Narrative theory and the cognitive sciences* Standford, CA: CSLI Publications.

Klein, S.B. (2013). The complex act of projecting oneself into the future. *Wiley Interdisciplinary*

*Reviews-Cognitive Science, 4*(1), pp. 63-79.

Klein, S.B., Cosmides, L., Tooby, J., et al. (2002). Decisions and the evolution of memory: multiple systems, multiple functions. *Psychological review, 109*(2), pp. 206-329.

Knauper, B., Roseman, M., Johnson, P.J., et al. (2009). Using Mental Imagery to Enhance the Effectiveness of Implementation Intentions. *Current Psychology, 28*(3), pp. 181-186.

Knott, A., & Sanders, T. (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics, 30*(2), pp. 135-175.

Kobayashi, C., Glover, G.H., & Temple, E. (2008). Switching language switches mind: linguistic effects on developmental neural bases of 'Theory of Mind'. *Social Cognitive and Affective Neuroscience, 3*(1), pp. 62-70.

Koch, C., & Hepp, K. (2006). Quantum mechanics in the brain. *Nature, 440*(7084), pp. 611-611.

Koechlin, E., & Jubault, T. (2006). Broca's Area and the Hierarchical Organization of Human Behavior. *Neuron, 50*(6), pp. 963-974.

Koelsch, S., Grossmann, T., Gunter, T.C., et al. (2003). Children processing music: Electric brain responses reveal musical competence and gender differences. *Journal of Cognitive Neuroscience, 15*(5), pp. 683-693.

Kolodny, N. (2008). The Myth of Practical Consistency. *European Journal of Philosophy, 16*(3), pp. 366-402.

Kording, K.P., & Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), pp. 244-247.

Koslicki, K. (2007). Towards a Neo-Aristotelian Mereology. *Dialectica, 61*(1), pp. 127-159.

Kosslyn, S.M. (2008). Visual Mental Imagery. In F.S. Kessel, P.L. Rosenfield & N.B. Anderson (Eds.), *Interdisciplinary research: case studies from health and social science* New York: Oxford University Press.

Kosslyn, S.M., & Maljkovic, V. (1990). Marr's Metatheory Revisited. *Concepts in Neuroscience, 1*(2), pp. 239-251.

Kriegel, U. (2006). Consciousness, Theories of. *Philosophy Compass, 1*(1), pp. 58-64.

Krueger, F., Barbey, A.K., & Grafman, J. (2009). The medial prefrontal cortex mediates social event knowledge. *Trends in Cognitive Sciences, 13*(3), pp. 103-109.

Kruglanski, A.W., & Orehek, E. (2007). Partitioning the Domain of Social Inference: Dual Mode and Systems Models and Their Alternatives. *Annual Review of Psychology, 58*(1), pp. 291-316.

Kunihiko, K., & Ichiro, T. (2003). Chaotic itinerancy. *Chaos, 13*(3), pp. 926-936.

Kurby, C.A., & Zacks, J.M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12*(2), pp. 72-79.

Lacey, S., Stilla, R., & Sathian, K. (2012). Metaphorically feeling: Comprehending textural

metaphors activates somatosensory cortex. *Brain and Language, 120*(3), pp. 416-421.

Lakatos, I. (1976). A Renaissance of Empiricism in the recent Philosophy of Mathematics? *The British journal for the philosophy of science, 27 (1976)*, pp. 201-223.

Lakoff, G. (2006). The Neuroscience of Form in Art. In M. Turner (Ed.), *The Artful Mind* Oxford: Oxford University Press.

Lamarque, P. (2004). On Not Expecting Too Much from Narrative. *Mind & Language, 19*(4), pp. 393-408.

Lamme, V.A.F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), pp. 494-501.

Lamme, V.A.F., & Roelfsema, P.R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences, 23*(11), pp. 571-579.

Lashley, K.S. (1951). The problem of serial order in behavior. In A.L. Jeffress (Ed.), *Cerebral mechanisms in behavior: the Hixon symposium* New York: Wiley, pp. 112-135.

Le, K., Coelho, C., Mozeiko, J., et al. (2011). Measuring Goodness of Story Narratives. *J Speech Lang Hear Res, 54*(1), pp. 118-126.

Lengfelder, A., & Gollwitzer, P.M. (2001). Reflective and reflexive action control in patients with frontal brain lesions. *Neuropsychology, 15*(1), pp. 80-100.

Leunissen, M. (2012). Aristotle on Natural Character and Its Implications for Moral Development. *Journal of the History of Philosophy, 50*(4), pp. 507-530.

Leuridan, B. (2010). Can Mechanisms Really Replace Laws of Nature? *Philosophy of Science, 77*(3), pp. 317-340.

Levitin, D.J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *NeuroImage, 20*(4), pp. 2142-2152.

Lhermitte, F. (1983). 'Utilization behaviour' and its relation to lesions of the frontal lobes. *Brain, 106*(2), pp. 237.

Lhermitte, F., Pillon, B., & Serdaru, M. (1986). Human autonomy and the frontal lobes. Part I: Imitation and utilization behavior: A neuropsychological study of 75 patients. *Annals of Neurology, 19*(4), pp. 326-334.

Liberman, N., Trope, Y., McCrea, S.M., et al. (2007). The effect of level of construal on the temporal distance of activity enactment. *Journal of Experimental Social Psychology, 43*(1), pp. 143-149.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences, 8*(4), pp. 529-539.

Libet, B. (2004). *Mind Time  The Temporal Factor in Consciousness* Cambridge MA: Harvard University Press.

Lieberman, M.D. (2007). Social Cognitive Neuroscience: A Review of Core Processes. *Annual*

*Review of Psychology, 58*(1), pp. 259-289.

Lieberman, M.D., Gaunt, R., Gilbert, D.T., et al. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology 34*, pp. 199-249.

Lloyd, G.E.R. (1982). *Early Greek Science: Thales to Aristotle* London: Chatto & Windus.

Lloyd, G.E.R. (2007). *Cognitive variations: reflections on the unity and diversity of the human mind*. Oxford: Oxford University Press.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), pp. 869-878.

Logothetis, N.K., & Schall, J.D. (1989). Neuronal correlates of subjective visual perception. *Science, 245*(4919), pp. 761-763.

Looren de Jong, H. (2002). Levels of explanation in biological psychology. *Philosophical Psychology, 15*(4), pp. 441-462.

Looren de Jong, H., & Schouten, M. (2007). Mind Reading and Mirror Neurons: Exploring Reduction. In M. Schouten & H. Looren de Jong (Eds.), *The Matter of the Mind. Philosophical Essays on Psychology, Neuroscience, and Reduction* Oxford: Blackwell Publishing, pp. 298-322.

Losin, E.A.R., Dapretto, M., & Iacoboni, M. (2010). Culture and neuroscience: additive or synergistic? *Social Cognitive and Affective Neuroscience, 5*(2-3), pp. 148-158.

Lovinger, D.M. (1993). Excitotoxicity and Alcohol-Related Brain Damage. *Alcoholism: Clinical and Experimental Research, 17*(1), pp. 19-27.

Lyons, D.E., Damrosch, D.H., Lin, J.K., et al. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1567), pp. 1158-1167.

Lyons, D.E., Young, A.G., & Keil, F.C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences of the United States of America, 104*(50), pp. 19751.

Machamer, P., Darden, L., & Craver, C.-F. (2000). Thinking about Mechanisms. *Philosophy of Science, 67*(1), pp. 1-25.

Mackenzie, C. (2008). Introduction. Practical identity and narrative agency. In C. Mackenzie & K. Atkins (Eds.), *Practical identity and narrative agency* New York: Routledge, pp. 1-28.

Macrae, C.N., & Johnston, L. (1998). Help, I Need Somebody: Automatic Action and Inaction. *Social Cognition, 16*(4), pp. 400-417.

Mahon, B.Z., Schwarzbach, J., & Caramazza, A. (2010). The Representation of Tools in Left Parietal Cortex Is Independent of Visual Experience. *Psychological Science, 21*(6), pp. 764-771.

Mak, W.M., & Sanders, T.J.M. (2012). The role of causality in discourse processing: Effects of

expectation and coherence relations. *Language and Cognitive Processes*, pp. 1-24.

Malafouris, L. (2010). The brain–artefact interface (BAI): a challenge for archaeology and cultural neuroscience. *Social Cognitive and Affective Neuroscience, 5*(2-3), pp. 264-273.

Mampe, B., Friederici, A.D., Christophe, A., et al. (2009). Newborns' Cry Melody Is Shaped by Their Native Language. *Current Biology, 19*(23), pp. 1994-1997.

Mar, R.A. (2004). The neuropsychology of narrative: story comprehension, story production and their interrelation. *Neuropsychologia, 42*(10), pp. 1414-1434.

Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences, 8*(2), pp. 79.

Marcus, G. (2009). How Does the Mind Work? Insights from Biology. *Topics in Cognitive Science, 1*(1), pp. 145-172.

Marcus, G.F. (2008). *Kluge: The haphazard construction of the human mind*. Boston, MA: Houghton Mifflin.

Mareschal, D., Johnson, M.H., Sirois, S., et al. (2007). *Neuroconstructivism: How the brain constructs cognition. Volume one*. Oxford: Oxford University Press.

Markman, A., & Dietrich, E. (2000). Extending the classical view of representation. *Trends in Cognitive Sciences, 4*(12), pp. 470-475.

Marr, D. (1976). Early Processing of Visual Information. *Philos Trans R Soc Lond B Biol Sci., 275*(942), pp. 485-519.

Marr, D. (1977a). Analysis of Occluding Contour. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 197*(1129), pp. 441-475.

Marr, D. (1977b). Artificial intelligence--A personal view. *Artificial Intelligence, 9*(1), pp. 37-48.

Marr, D. (1980). Visual information processing: the structure and creation of visual representations. *Philos Trans R Soc Lond B Biol Sci., 290*(1038), pp. 199-218.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Marr, D., & Hildreth, E. (1980). Theory of Edge Detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 207*(1167), pp. 187-217.

Marr, D., & Nishihara, H.K. (1978). Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 200*(1140), pp. 269-294.

Marr, D., & Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin, 15*(3), pp. 470-488.

Marr, D., & Poggio, T. (1979). A Computational Theory of Human Stereo Vision. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 204*(1156), pp. 301-328.

Masson, M.E.J., Bub, D.N., & Breuer, A.T. (2011). Priming of reach and grasp actions by

handled objects. *Journal of Experimental Psychology: Human Perception and Performance, 37*(5), pp. 1470-1484.

Mathur, V.A., Harada, T., & Chiao, J.Y. (2012). Racial identification modulates default network activity for same and other races. *Human Brain Mapping, 33*(8), pp. 1883-1893.

Mayr, E. (1964). The evolution of living systems. *Proceedings of the National Academy of Sciences of the United States of America, 51*(5), pp. 934-941.

Mayr, E. (1974). Behavior programs and evolutionary strategies. *American scientist, 62*(6), pp. 650-659.

McCauley, R.N. (1986). Intertheoretic Relations and the Future of Psychology. *Philosophy of Science, 53*(2), pp. 179-199.

McCauley, R.N., & Bechtel, W. (2001). Explanatory Pluralism and Heuristic Identity Theory. *Theory & Psychology, 11*(6), pp. 736-760.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines, 1*(2), pp. 185-196.

McCulloch, W. (1945). A heterarchy of values determined by the topology of nervous nets. *Bulletin of Mathematical Biology, 7*(2), pp. 89-93.

McDermott, J., & Hauser, M.D. (2007). Nonhuman primates prefer slow tempos but dislike music overall. *Cognition, 104*(3), pp. 654-668.

McDowell, J.H. (1994). *Mind and world*. Cambridge, MA: Harvard University Press.

McShea, D.W. (1991). Complexity and evolution: What everybody knows. *Biology and Philosophy, 6*(3), pp. 303-324.

McVee, M.B., Dunsmore, K., & Gavelek, J.R. (2005). Schema Theory Revisited. *Review of Educational Research, 75*(4), pp. 531-566.

Meeter, M., Jehee, J., & Murre, J. (2007). Neural Models that Convince: Model Hierarchies and Other Strategies to Bridge the Gap Between Behavior and the Brain. *Philosophical Psychology, 20*(6), pp. 749 - 772.

Meinz, E.J., & Hambrick, D.Z. (2010). Deliberate Practice Is Necessary but Not Sufficient to Explain Individual Differences in Piano Sight-Reading Skill. *Psychological Science, 21*(7), pp. 914-919.

Mele, A.R. (1992). *Springs of action*. New York: Oxford University Press.

Mele, A.R., & Moser, P.K. (1994). Intentional Action. *Noûs, 28*(1), pp. 39-68.

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture, 4*(1), pp. 1.

Meteyard, L., Cuadrado, S.R., Bahrami, B., et al. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), pp. 788-804.

Metzinger, T. (2000). Introduction: Consciousness Research at the End of the Twentieth century. In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual*

*questions.* Cambridge, MA: MIT Press, pp. 1-12.

Meunier, D., Lambiotte, R., & Bullmore, E.T. (2010). Modular and hierarchically modular organization of brain networks. *Frontiers in Neuroscience, 4* pp. 1-11.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), pp. 81.

Miller, G.A., Galanter, E., & Pribram, K.H. (1960). Motor skills and habits. In G.A. Miller, E. Galanter & K.H. Pribram (Eds.), *Plans and the structure of behavior* New York: Henry Holt, pp. 81-93.

Millgram, E. (2000). Coherence: The Price of the Ticket. *The Journal of Philosophy, 97*(2), pp. 82-93.

Milne, E., & Grafman, J. (2001). Ventromedial Prefrontal Cortex Lesions in Humans Eliminate Implicit Gender Stereotyping. *The Journal of Neuroscience, 21*(12), pp. 1-6.

Minshew, N.J., Williams, D.L., & McFadden, K. (2009). Information Processing, Neural Connectivity, and Neuronal Organization. *Autism*, pp. 381-405.

Minsky, M. (1975). A framework for representing knowledge. In P.H. Winston (Ed.), *The Psychology of Computer Vision* New York: McGraw-Hill, pp. 211-281.

Missios, S. (2007). Hippocrates, Galen, and the uses of trepanation in the ancient classical world. *Neurosurgical Focus, 23*(1), pp. E11.

Mitchell, S.D. (2002). Integrative Pluralism. *Biology and Philosophy, 17*(1), pp. 55-70.

Mitchell, S.D. (2006). Modularity - More Than a Buzzword? *Biological Theory, 1*(1), pp. 98-101.

Mitchell, S.D., & Dietrich, M.R. (2006). Integration without Unification: An Argument for Pluralism in the Biological Sciences. *The American Naturalist, 168*(S6), pp. S73-S79.

Mithen, S.J. (2005). *The singing Neanderthals: the origins of music, language, mind, and body* London: Weidenfeld & Nicolson.

Molenaar, P.C.M. (2001). Review: 'Conceptual Spaces: The Geometry of Thought; Peter Gärdenfors' *Acta Psychologica, 106*(3), pp. 333-336.

Moll, J., Zahn, R., de Oliveira-Souza, R., et al. (2005). The neural basis of human moral cognition. *Nat Rev Neurosci, 6*(10), pp. 799-809.

Monti, M.M., Vanhaudenhuyse, A., Coleman, M.R., et al. (2010). Willful Modulation of Brain Activity in Disorders of Consciousness. *New England Journal of Medicine, 362*(7), pp. 579-589.

Moors, A., & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin, 132*(2), pp. 297-326.

Morishima, Y., Schunk, D., Bruhin, A., et al. (2012). Linking Brain Structure and Activation in Temporoparietal Junction to Explain the Neurobiology of Human Altruism. *Neuron, 75*(1), pp. 73-79.

Morton, P. (1993). Supervenience and Computational Explanation in Vision Theory.

*Philosophy of Science, 60*(1), pp. 86-99.

Moss, L. (2012). Is The Philosophy of Mechanism Philosophy Enough? *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), pp. 164-172.

Mozeiko, J., Le, K., Coelho, C., et al. (2011). The relationship of story grammar and executive function following TBI. *Aphasiology, 25*(6-7), pp. 826-835.

Munsell, A.H. (1912). A Pigment Color System and Notation. *The American journal of psychology, 23*(2), pp. 236-244.

Murphy, J.B. (2002). Nature, Custom, and Reason as the Explanatory and Practical Principles of Aristotelian Political Science. *The Review of Politics, 64*(3), pp. 469-495.

Mushiake, H., Saito, N., Sakamoto, K., et al. (2006). Activity in the Lateral Prefrontal Cortex Reflects Multiple Steps of Future Events in Action Plans. *Neuron, 50*(4), pp. 631-641.

Nachev, P., & Hacker, P.M.S. (2010). Covert cognition in the persistent vegetative state. *Progress in Neurobiology, 91*(1), pp. 68-76.

Nadel, L. (1992). Multiple Memory Systems: What and Why. *Journal of Cognitive Neuroscience, 4*(3), pp. 179-188.

Nakao, T., Ohira, H., & Northoff, G. (2012). Distinction between externally vs. internally guided decision-making: Operational differences, meta-analytical comparisons and their theoretical implications. *Frontiers in Neuroscience, 6*, pp. 1-26.

Nakata, T., & Trehub, S.E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development, 27*(4), pp. 455-464.

Narayanan, S. (2009). *Mind changes: A simulation semantics account of counterfactuals*. ICSI and UC Berkeley. Berkeley, CA. Retrieved on August 30, 2013 from http://www.icsi.berkeley.edu/~snarayan/counterfactuals.pdf

Nathaniel-James, D.A., & Frith, C.D. (2002). The Role of the Dorsolateral Prefrontal Cortex: Evidence from the Effects of Contextual Constraint in a Sentence Completion Task. *NeuroImage, 16*(4), pp. 1094-1102.

Neisser, J. (2012). Neural correlates of consciousness reconsidered. *Consciousness and Cognition, 21*(2), pp. 681-690.

Nelson, K. (1999). Event representations, narrative development and internal working models. *Attachment & Human Development, 1*(3), pp. 239-252.

Nelson, K. (2003). Self and social functions: Individual autobiographical memory and collective narrative. *Memory, 11*(2), pp. 125-136.

Nelson, K. (2007). Developing past and future selves for time travel narratives. *Behavioral and Brain Sciences, 30*(03), pp. 327-328.

Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.

Newsome, M.R., Scheibel, R.S., Hanten, G., et al. (2010). Brain Activation While Thinking About the Self From Another Person's Perspective After Traumatic Brain Injury in

Adolescents. *Neuropsychology, 24*(2), pp. 139-147.

Ng, J., Bharath, A.A., & Zhaoping, L. (2007). A survey of architecture and function of the primary visual cortex (V1). *EURASIP Journal on Applied Signal Processing, 2007*(1), pp. 1-17.

Niedenthal, P.M., Barsalou, L.W., Winkielman, P., et al. (2005). Embodiment in Attitudes, Social Perception, and Emotion. *Personality and Social Psychology Review, 9*(3), pp. 184-211.

Nielsen, K.S. (2010). Representation and dynamics. *Philosophical Psychology, 23*(6), pp. 759 - 773.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in Kalahari Bushman Children and the Origins of Human Cultural Cognition. *Psychological Science, 21*(5), pp. 729-736.

Nisbett, R.E., & Miyamoto, Y. (2005). The influence of culture: holistic versus analytic perception. *Trends in Cognitive Sciences, 9*(10), pp. 467-473.

Nisbett, R.E., Peng, K., Choi, I., et al. (2001). Culture and Systems of Thought: Holistic Versus Analytic Cognition. *Psychological Review, 108*(2), pp. 291-310.

Nishitani, N., Schurmann, M., Amunts, K., et al. (2005). Broca's Region: From Action to Language. *Physiology, 20*(1), pp. 60-69.

Noë, A. (2004). *Action in Perception*. Cambridge, MA: MIT Press.

Noë, A., & Thompson, E. (2004a). Are There Neural Correlates of Consciousness? *Journal of Consciousness Studies, 11*, pp. 3-28.

Noë, A., & Thompson, E. (2004b). Sorting Out the Neural Basis of Consciousness Authors' Reply to Commentators. *Journal of Consciousness Studies, 11*, pp. 87-98.

Norman, D., & Shallice, T. (1986). Attention to action: willed and automatic control of behavior. In R. Davidson, R. Schwartz & D. Shapiro (Eds.), *Consciousness and Self-Regulation: Advances in Research and Theory IV* New York: Plenum Press.

Northoff, G. (2004). *Philosophy of the Brain: The" Brain Problem* (Vol. 52). Amsterdam: John Benjamins.

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences, 8*(3), pp. 102-107.

Nudo, R., Milliken, G., Jenkins, W., et al. (1996). Use-dependent alterations of movement representations in primary motor cortex of adult squirrel monkeys. *The Journal of Neuroscience, 16*(2), pp. 785-807.

O'Doherty, J.E., Lebedev, M.A., Ifft, P.J., et al. (2011). Active tactile exploration using a brain-machine-brain interface. *Nature, 479*, pp. 228-231.

O'Reilly, R.C., & Rudy, J.W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review, 108*(2), pp. 311-345.

Ochsner, K.N., & Kosslyn, S.M. (1999). The Cognitive Neuroscience Approach. In B.M. Bly & D.E. Rumelhart (Eds.), *Cognitive Science* San Diego, CA: Academic Press.

Oele, M. (2007). Being Beyond: Aristotle's and Plessner's Accounts of Animal Responsiveness.

In C. Painter & C. Lotz (Eds.), *Phenomenology And The Non-Human Animal* Dordrecht: Springer, pp. 29-37.

Okuda, J., Gilbert, S.J., Burgess, P.W., et al. (2011). Looking to the future: Automatic regulation of attention between current performance and future plans. *Neuropsychologia, 49*(8), pp. 2258-2271.

Orden, G.C.v., Pennington, B.F., & Stone, G.O. (2001). What do double dissociations prove? *Cognitive Science: A Multidisciplinary Journal, 25*(1), pp. 111-172.

Osvath, M. (2009). Spontaneous planning for future stone throwing by a male chimpanzee. *19*(5), pp. R190-R191.

Pacherie, E. (2000). The content of intentions. *Mind and Language, 15*(4), pp. 400-432.

Pacherie, E. (2001). Agency lost and found: a commentary on Spence. *Philosophy, Psychiatry, & Psychology, 8*(2), pp. 173-176.

Pacherie, E. (2006). Towards a dynamic theory of intentions. In S. Pockett, W.P. Banks & S. Gallagher (Eds.), *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition.* Cambridge, MA: MIT Press, pp. 145-167.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*(1), pp. 179-217.

Pacherie, E. (2011). Nonconceptual Representations for Action and the Limits of Intentional Control. *Social Psychology, 42*(1), pp. 67-73.

Pacherie, E., & Haggard, P. (2010). What are Intentions? In L. Nadel & W. Sinnott-Armstrong (Eds.), *Conscious Will and Responsibility: A Tribute to Benjamin Libet* Oxford: Oxford University Press.

Painter, C., & Lotz, C. (2007). Introduction. Phenomenology And The Non-Human Animal. In C. Painter & C. Lotz (Eds.), *Phenomenology and the Non-Human Animal* Dordrecht: Springer, pp. 1-11.

Panksepp, J., & Northoff, G. (2009). The trans-species core SELF: The emergence of active cultural and neuro-ecological agents through self-related processing within subcortical-cortical midline networks. *Consciousness and Cognition, 18*(1), pp. 193-215.

Panzer, S., Krueger, M., Muehlbauer, T., et al. (2009). Inter-manual transfer and practice: Coding of simple motor sequences. *Acta Psychologica, 131*(2), pp. 99-109.

Papies, E.K., Aarts, H., & de Vries, N.K. (2009). Planning is for doing: Implementation intentions go beyond the mere creation of goal-directed associations. *Journal of Experimental Social Psychology, 45*(5), pp. 1148-1151.

Papineau, D. (2006). *The Roots of Reason: Philosophical Essays on Rationality, Evolution, and Probability: Philosophical Essays on Rationality, Evolution, and Probability* Oxford: Oxford University Press

Papousek, H. (1996). Musicality in infancy research: biological and cultural origins of early

musicality. In I. Deliège & J.A. Sloboda (Eds.), *Musical beginnings: Origins and development of musical competence* Oxford: Oxford University Press, pp. 37-55.

Park, D.C., & Huang, C.-M. (2010). Culture Wires the Brain. *Perspectives on Psychological Science, 5*(4), pp. 391-400.

Parry, M. (1930). Studies in the Epic Technique of Oral Verse-Making. I. Homer and Homeric Style. *Harvard Studies in Classical Philology, 41*, pp. 73-147.

Peacocke, C. (1986). Explanation in Computational Psychology: Language, Perception and Level 1.5. *Mind & Language, 1*(2), pp. 101-123.

Pecher, D., Boot, I., & Van Dantzig, S. (2011). Abstract Concepts: Sensory-Motor Grounding, Metaphors, and Beyond. In H.R. Brian (Ed.), *Psychology of Learning and Motivation* Vol. Volume 54: Academic Press, pp. 217-248.

Peeters, R., Simone, L., Nelissen, K., et al. (2009). The Representation of Tool Use in Humans and Monkeys: Common and Uniquely Human Features. *J. Neurosci., 29*(37), pp. 11523-11539.

Pellegrin, P. (1987). Logical difference and biological difference: the unity of Aristotle's thought. In A. Gotthelf & J.G. Lennox (Eds.), *Philosophical Issues in Aristotle's Biology* Cambridge: Cambridge University Press, pp. 313-338.

Pennartz, C.M.A., Uylings, H.B.M., Barnes, C.A., et al. (2002). Memory reactivation and consolidation during sleep: from cellular mechanisms to human performance. In M.A. Hofman, G.J. Boer, A.J.G.D. Holtmaat, E.J.W. Van Someren, J. Verhaagen & D.F. Swaab (Eds.), *Progress in Brain Research* Vol. Volume 138: Elsevier, pp. 143-166.

Pennebaker, J.W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy, 31*(6), pp. 539-548.

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition, 100*(1), pp. 1-32.

Peretz, I., & Zatorre, R.J. (2005). Brain Organization for Music Processing. *Annual Review of Psychology, 56*(1), pp. 89-114.

Petersen, S.E., van Mier, H., Fiez, J.A., et al. (1998). The effects of practice on the functional anatomy of task performance. *PNAS, 95*(3), pp. 853-860.

Pezzulo, G., Barsalou, L.W., Cangelosi, A., et al. (2011). The mechanics of embodiment: a dialog on embodiment and computational modeling. *Frontiers in Psychology, 2*, pp. 5-5.

Pezzulo, G., & Ognibene, D. (2012). Proactive Action Preparation: Seeing Action Preparation as a Continuous and Proactive Process. *Motor Control, 16*(3), pp. 386-424.

Phillips, J.K., Klein, G., & Sieck, W.R. (2008). Expertise in Judgment and Decision Making: A Case for Training Intuitive Decision Skills. In D.J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* Malden, MA: Blackwell Publishing, pp. 297-315.

Piccini, G., & Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analyses

as Mechanism Sketches. *Synthese 183*(3), pp. 283-311.

Piccinini, G. (2008). Computation without Representation. *Philosophical Stud., 137*(2), pp. 205-241.

Piccinini, G., & Scarantino, A. (2010). Computation vs. information processing: why their difference matters to cognitive science. *Studies In History and Philosophy of Science Part A, 41*(3), pp. 237-246.

Plaut, D.C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*(2), pp. 291 - 321.

Plautz, E.J., Milliken, G.W., & Nudo, R.J. (2000). Effects of Repetitive Motor Training on Movement Representations in Adult Squirrel Monkeys: Role of Use versus Learning. *Neurobiology of Learning and Memory, 74*(1), pp. 27-55.

Poggio, T. (1981). Marr's computational approach to vision. *Trends in Neurosciences, 4*, pp. 258-262.

Poggio, T. (2010). Afterword. Marr's Vision and Computational Neuroscience *Vision : a computational investigation into the human representation and processing of visual information* Cambridge, MA: MIT press.

Poldrack, R.A., Sabb, F.W., Foerde, K., et al. (2005). The Neural Correlates of Motor Skill Automaticity. *J. Neurosci., 25*(22), pp. 5356-5364.

Polger, T.W. (2004). Neural machinery and realization. *Philosophy Of Science, 71*(5), pp. 997-1006.

Pollard, B. (2003). Can Virtuous Actions be Both Habitual and Rational? *Ethical Theory and Moral Practice, 6*(4), pp. 411-425.

Pollard, B. (2005). Naturalizing the space of reasons. *Int. J. of Philosophical Studies, 13*(1), pp. 69-82.

Pollard, B. (2006). Explaining Actions with Habits. *Amer. Philosophical Quarterly, 43*(1), pp. 57-69.

Povinelli, D.J., & Barth, J. (2005). Reinterpreting behavior: A human specialization? *Behavioral and Brain Sciences, 28*(05), pp. 712-713.

Preston, A.R., Shrager, Y., Dudukovic, N.M., et al. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14*(2), pp. 148-152.

Price, D. de Solla (1974). Gears from the Greeks. The Antikythera Mechanism: A Calendar Computer from ca. 80 B. C. *Transactions of the American Philosophical Society, 64*(7), pp. 1-70.

Prinz, J. (2006). Is the Mind Really Modular? In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science* Malden, MA: Blackwell Publishing.

Propp, V.I. (2003). *Morphology of the folk tale* (L. Scott, Trans.). Austin, TX: University of Texas Press.

Pulvermüller, F. (2012). Meaning and the brain: The neurosemantics of referential, interactive, and combinatorial knowledge. *Journal of Neurolinguistics, 25*(5), pp. 423-459.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nat Rev Neurosci, 11*(5), pp. 351-360.

Pulvermüller, F., Hauk, O., Nikulin, V.V., et al. (2005). Functional links between motor and language systems. *European Journal of Neuroscience, 21*(3), pp. 793-797.

Putnam, H. (1967). Psychological predicates. In W.H. Capitan & D.D. Merrill (Eds.), *Art, Mind and Religion* Pittsburgh: University of Pittsburgh Press, pp. 37–48.

Putnam, H. (1975). *Mind, Language and Reality*: Cambridge University Press.

Quante, M. (2008). Review of: P.M.S. Hacker: Human Nature: The Categorial Framework. *Notre Dame Philosophical Reviews, 6*(41). Retrieved on August 30, 2013 from http://ndpr.nd.edu/review.cfm?id=13430

Quartz, S.R. (1999). The constructivist brain. *Trends in Cognitive Sciences, 3*(2), pp. 48-57.

Raffard, S., D'Argembeau, A., Bayard, S., et al. (2010). Scene Construction in Schizophrenia. *Neuropsychology, 24*(5), pp. 608-615.

Raichle, M.E. (1998). The neural correlates of consciousness: an analysis of cognitive skill learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 353*(1377), pp. 1889.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., et al. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), pp. 676-682.

Raichle, M.E., & Snyder, A.Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage, 37*(4), pp. 1083-1090.

Raijmakers, M.E.J. (2007). Modelling cognitive developmental transitions in neural networks: bifurcations in an adaptive resonance theory model. In D. Mareschal, S. Sirois, M.H. Johnson & G. Westermann (Eds.), *Neuroconstructivism: Perspectives and Prospects* Oxford: Oxford University Press, pp. 99-128.

Raposo, A., Moss, H.E., Stamatakis, E.A., et al. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia, 47*(2), pp. 388-396.

Rasmussen, N. (1987). A new model of developmental constraints as applied to the Drosophila system. *Journal of Theoretical Biology, 127*(3), pp. 271-299.

Reber, T.P., Luechinger, R., Boesiger, P., et al. (2012). Unconscious Relational Inference Recruits the Hippocampus. *The Journal of Neuroscience, 32*(18), pp. 6138-6148.

Redgrave, P., Prescott, T.J., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the

selection problem? *Neuroscience, 89*(4), pp. 1009-1023.

Reed, N., McLeod, P., & Dienes, Z. (2010). Implicit knowledge and motor skill: What people who know how to catch don't know. *Consciousness and Cognition, 19*(1), pp. 63-76.

Rees, G., & Frith, C. (2007a). A brief history of the scientific approach to the study of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness.* Malden, MA: Blackwell, pp. 9-26.

Rees, G., & Frith, C. (2007b). Methodologies for identifying the neural correlates of consciousness. In M. Velmans & S. Schneider (Eds.), *The Blackwell companion to consciousness* Malden, MA: Blackwell, pp. 453-466.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences, 104*(4), pp. 1436-1441.

Revonsuo, A. (2000). Prospects for a scientific research program on consciousness. In T. Metzinger (Ed.), *Neural correlates of consciousness* Cambridge, MA: MIT Press, pp. 57-75.

Richmond, J.L., & Pan, R. (2013). Thinking about the future early in life: The role of relational memory. *Journal of Experimental Child Psychology, 114*(4), pp. 510-521.

Ricoeur, P. (1970). *Freud and philosophy. An essay on interpretation* (D. Savage, Trans.). New Haven: Yale University Press.

Ricoeur, P. (1971). The Model of the Text: Meaningful Action Considered as a Text. *Social research, 38*(3), pp. 185-218.

Ricoeur, P. (1980). Narrative Time. *Critical Inquiry, 7*(1), pp. 169-190.

Ricoeur, P. (1984-88). *Time and narrative (vols. 1-3)* (K. McLaughlin & D. Pellauer, Trans.). Chicago: University of Chicago Press.

Ricoeur, P. (1984). *Time and narrative. (volume 1)* (K. McLaughlin & D. Pellauer, Trans.). Chicago: University of Chicago Press.

Ricoeur, P. (1985). *Time and narrative. (volume 2)* (K. McLaughlin & D. Pellauer, Trans.). Chicago: University of Chicago Press.

Ricoeur, P. (1988). *Time and narrative. (volume 3)* (K. Blamey & D. Pellauer, Trans.). Chicago: University of Chicago Press.

Ricoeur, P. (1991). Narrative identity. *Philosophy Today, 35*(1), pp. 73-81.

Ricoeur, P. (1991a). Life in Quest of Narrative. In D.E. Wood (Ed.), *On Paul Ricoeur: Narrative and Interpretation,* New York: Routledge, pp. 20-33.

Ricoeur, P. (1991b). Life: a Story in Search of a Narrator. In M.J. Valdés (Ed.), *A Ricoeur reader: reflection and imagination*, New York: Harvester Wheatsheaf.

Ricoeur, P. (1991c). Mimesis and Representation. In M.J. Valdés (Ed.), *Reflection and imagination.* New York: Harvester Wheatsheaf.

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). Chicago: University of Chicago Press.

Ridderinkhof, K.R., van den Wildenberg, W.P.M., Segalowitz, S.J., et al. (2004). Neurocognitive mechanisms of cognitive control: the role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition, 56*(2), pp. 129-140.

Rietveld, E. (2008). *Unreflective action. A philosophical contribution to integrative neuroscience*. Amsterdam: Institute for Logic, Language and Computation.

Rizzolatti, G., & Arbib, M.A. (1998). Language within our grasp. *Trends in Neurosciences, 21*(5), pp. 188-194.

Rizzolatti, G., Fadiga, L., Gallese, V., et al. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*(2), pp. 131-141.

Rizzolatti, G., & Sinigaglia, C. (2008). *Mirrors in the Brain: How Our Minds Share Actions and Emotions*: Oxford University Press.

Roberson, D., Hanley, J.R., & Pak, H. Thresholds for color discrimination in English and Korean speakers. *Cognition 11*(3), pp. 482-487.

Roepstorff, A. (2008). Things to think with: words and objects as material symbols. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1499), pp. 2049-2054.

Roepstorff, A. (2013). Why Am I Not Just Lovin' Cultural Neuroscience? Toward a Slow Science of Cultural Difference. *Psychological Inquiry, 24*(1), pp. 61-63.

Roepstorff, A., & Frith, C. (2004). What's at the top in the top-down control of action? Script-sharing and 'top-top' control of action in cognitive experiments. *Psychological Research, 68*(2), pp. 189-198.

Roepstorff, A., & Frith, C. (2012). Neuroanthropology or simply anthropology? Going experimental as method, as object of study, and as research aesthetic. *Anthropological Theory, 12*(1), pp. 101-111.

Rolls, E.T. (2011). David Marr's Vision: floreat computational neuroscience. *Brain, 134*(3), pp. 913-916.

Romero, K., & Moscovitch, M. (2012). Episodic memory and event construction in aging and amnesia. *Journal of Memory and Language, 67*(2), pp. 270-284.

Rosenbaum, D.A., Carlson, R.A., & Gilmore, R.O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology, 52*(1), pp. 453-470.

Ross, E.D. (2010). Cerebral Localization of Functions and the Neurology of Language: Fact versus Fiction or Is It Something Else? *Neuroscientist, 16*(3), pp. 222-243.

Rossetti, Y. (1998). Implicit Short-Lived Motor Representations of Space in Brain Damaged and Healthy Subjects. *Consciousness and Cognition, 7*(3), pp. 520-558.

Rouw, R., & Scholte, H.S. (2007). Increased structural connectivity in grapheme-color synesthesia. *Nat Neurosci, 10*(6), pp. 792-797.

Rowlands, M. (2012). Representing Without Representations. *AVANT, III*(1), pp. 133-144.

Ruben, D.H. (1992). *Explaining Explanation*. Routledge: New York 1992.

Rumiati, R.I., Zanini, S., Vorano, L., et al. (2001). A Form of Ideational Apraxia as a Delective Deficit of Contention Scheduling. *Cognitive Neuropsychology, 18*(7), pp. 617-642.

Rupert, R.D. (2009). *Cognitive systems and the extended mind*. New York: Oxford University Press.

Rupert, R.D. (2010). Extended cognition and the priority of cognitive systems. *Cognitive Systems Research, 11*(4), pp. 343-356.

Rydell, R.J., McConnell, A.R., Mackie, D.M., et al. (2006). Of Two Minds: Forming and Changing Valence-Inconsistent Implicit and Explicit Attitudes. *Psychological Science, 17*, pp. 954-958.

Schack, T., & Hackfort, D. (2007). Action-Theory Approach to Applied Sport Psychology. In G. Tenenbaum & R.C. Eklund (Eds.), *Handbook of Sport Psychology* Hoboken, N.J.: John Wiley, pp. 332-351.

Schack, T., & Mechsner, F. (2006). Representation of motor skills in human long-term memory. *Neuroscience Letters, 391*(3), pp. 77-81.

Schacter, D.L., & Addis, D.R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), pp. 773-786.

Schacter, D.L., & Addis, D.R. (2007a). Constructive memory: The ghosts of past and future. *Nature, 445*(7123), pp. 27-27.

Schacter, D.L., & Addis, D.R. (2007b). On the constructive episodic simulation of past and future events. *Behavioral and Brain Sciences, 30*(03), pp. 331-332.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2007). Remembering the past to imagine the future: the prospective brain. *Nat Rev Neurosci, 8*(9), pp. 657-661.

Schacter, D.L., Addis, D.R., & Buckner, R.L. (2008). Episodic Simulation of Future Events. Concepts, data, and applications. *Annals of the New York Academy of Sciences, 1124*(The Year in Cognitive Neuroscience 2008), pp. 39-60.

Schaffer, J. (2003). Is There a Fundamental Level? *Nous, 37*(3), pp. 498-517.

Schaffer, J. (2005). Contrastive Causation. *The Philosophical Review, 114*(3), pp. 327-358.

Schank, J.C., & Wimsatt, W.C. (1986). Generative Entrenchment and Evolution. *PSA: Proc. of the Biennial Meeting of the Phil. of Science Ass., 1986*, pp. 33-60.

Schank, R.C. (1980). Language and memory. *Cognitive Science, 4*(3), pp. 243-284.

Schank, R.C., & Abelson, R.P. (1977). *Scripts, plans, goals and understanding. An inquiry into human knowledge structures*. Hillsdale, NJ: Laurence Erlbaum.

Schechtman, M. (2011). The narrative self. In S. Gallagher (Ed.), *The Oxford Handbook of the Self* Oxford: Oxford University Press, pp. 394-416.

Schmidt, R.A. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82*(4), pp. 225-260.

Schnakers, C., Perrin, F., Schabus, M., et al. (2008). Voluntary brain processing in disorders of consciousness. *Neurology, 71*(20), pp. 1614-1620.

Schneider, F., Bermpohl, F., Heinzel, A., et al. (2008). The Resting Brain and Our Self: Self-Relatedness Modulates Resting State Neural Activity in Cortical Midline Structures. *Neuroscience, 157*(1), pp. 120-131.

Schwartz, M.F. (2006). The cognitive neuropsychology of everyday action and planning. *Cognitive Neuropsychology, 23*(1), pp. 202-221.

Schwartz, M.F., Montgomery, M.W., Buxbaum, L.J., et al. (1998). Naturalistic action impairment in closed head injury. *Neuropsychology, 12*(1), pp. 13-28.

Schwartz, M.F., Montgomery, M.W., Fitzpatrick-desalme, E.J., et al. (1995). Analysis of a disorder of everyday action. *Cognitive Neuropsychology, 12*(8), pp. 863-892.

Schwartz, M.F., Reed, E.S., Montgomery, M., et al. (1991). The Quantitative Description of Action Disorganisation after Brain Damage: A Case Study. *Cognitive Neuropsychology, 8*(5), pp. 381-414.

Scott, W.A. (1962). Cognitive Complexity and Cognitive Flexibility. *Sociometry, 25*(4), pp. 405-414.

Scott-Baumann, A. (2009). *Ricoeur and the Hermeneutics of Suspicion* London: Continuum.

Searle, J.R. (1980). The intentionality of intention and action. *Cognitive Science, 4*(1), pp. 47-70.

Searle, J.R. (1983). *Intentionality. An essay in the philosophy of mind*. Cambridge: Cambridge U Press.

Searle, J.R. (2001). *Rationality in action* Cambridge, Mass: MIT Press |c c2001.

Searle, J.R. (2004). Comments on Noë and Thompson, 'Are There Neural Correlates of Consciousness?'. *Journal of Consciousness Studies, 11*(1), pp. 80-82.

Segalowitz, S.J. (2009). A quantum physics account of consciousness: Much less than meets the eye. *Brain and Cognition, 71*(2), pp. 53-53.

Seligman, R., & Kirmayer, L. (2008). Dissociative Experience and Cultural Neuroscience: Narrative, Metaphor and Mechanism. *Culture, Medicine and Psychiatry, 32*(1), pp. 31-64.

Selinger, E.M., & Crease, R.P. (2002). Dreyfus on expertise: The limits of phenomenological analysis. *Continental Philosophy Review, 35*(3), pp. 245-279.

Sellars, W. (1997). *Empiricism and the Philosophy of Mind*. Cambridge, MA: Cambridge U Press.

Seok, B. (2006). Diversity and Unity of Modularity. *Cognitive Science, 30*(2), pp. 347 - 380.

Seth, A.K., Izhikevich, E., Reeke, G.N., et al. (2006). Theories and measures of consciousness: An extended framework. *PNAS, 103*(28), pp. 10799-10804.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Shallice, T. (2002). Fractionation of the supervisory system. *Principles of frontal lobe function*, pp. 261-277.

Shallice, T., & Burgess, P. (1996). The Domain of Supervisory Processes and Temporal Organization of Behaviour. *Philosophical Transactions: Biological Sciences, 351*(1346), pp. 1405-1412.

Shanahan, M., & Baars, B. (2005). Applying global workspace theory to the frame problem. *Cognition, 98*(2), pp. 157.

Shea, N. (2012, in press). Reward prediction error signals are meta-representational. *Noûs*, pp. 1-31.

Sheeran, P., Webb, T.L., & Gollwitzer, P.M. (2005). The Interplay Between Goal Intentions and Implementation Intentions. *Personality and Social Psychology Bulletin, 31*(1), pp. 87-98.

Sheng, F., & Han, S. (2012). Manipulations of cognitive strategies and intergroup relationships reduce the racial bias in empathic neural responses. *NeuroImage, 61*(4), pp. 786-797.

Sherman, J.W., Gawronski, B., Gonsalkorale, K., et al. (2008). The Self-Regulation of Automatic Associations and Behavioral Impulses. *Psychological Review, 115*(2), pp. 314-335.

Sherman, N. (1989). *The Fabric of Character: Aristotle's Theory of Virtue*: Oxford University Press.

Sherry, D.F., & Schacter, D.L. (1987). The evolution of multiple memory-systems. *Psychological Review, 94*(4), pp. 439-454.

Shore, B. (1996). *Culture in Mind. Cognition, Culture, and the Problem of Meaning*. Oxford: Oxford University Press.

Silverberg, A. (2006). Chomsky and Egan on computational theories of vision. *Minds and Machines, 16*(4), pp. 495.

Simmons, W.K., & Barsalou, L.W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20*(3-6), pp. 451-486.

Simon, H.A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society, 106*(6), pp. 467-482.

Simon, H.A. (1973). The organization of complex systems. In H.H. Pattee (Ed.), *Hierarchy theory: The challenge of complex systems* New York: George Braziller, pp. 1-27.

Sirigu, A., Zalla, T., Pillon, B., et al. (1996). Encoding of sequence and boundaries of scripts following prefrontal lesions. *Cortex, 32*, pp. 297-310.

Sirigu, A., Zalla, T., Pillon, B., et al. (1995). Planning and Script Analysis following Prefrontal Lobe Lesions. *Annals of the New York Academy of Sciences, 769*(1), pp. 277-288.

Sirois, S., Spratling, M., Thomas, M.S.C., et al. (2008). Precis of Neuroconstructivism: How the Brain Constructs Cognition. *Behavioral and Brain Sciences, 31*(03), pp. 321-331.

Smallwood, J., Brown, K., Baird, B., et al. (2012). Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought.

*Brain Research, 1428*(0), pp. 60-70.

Smallwood, J., Tipper, C., Brown, K., et al. (2013). Escaping the here and now: Evidence for a role of the default mode network in perceptually decoupled thought. *NeuroImage, 69*(0), pp. 120-125.

Smith, C.U.M. (2009). The `hard problem' and the quantum physicists. Part 2: Modern times. *Brain and Cognition, 71*(2), pp. 54-63.

Smith, E.R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems. *Personality and Social Psychology Review, 4*(2), pp. 108-131.

Snell, B. (1928). *Aischylos und das Handeln im Drama* Leipzig: Dieterich.

Snell, B. (1975). *Die Entdeckung des Geistes*. Göttingen: Vandenhoeck & Ruprecht.

Sorabji, R. (1980). *Necessity, cause and blame: perspectives on Aristotle's theory*. London: Duckworth.

Sperber, D. (1985). Anthropology and Psychology: Towards an Epidemiology of Representations. *Man, 20*(1), pp. 73-89.

Sperber, D. (1996). *Explaining culture: a naturalistic approach*. Oxford: Blackwell.

Sperber, D. (2005). Modularity and relevance: How can a massively modular mind be flexible and context-sensitive? In P. Carruthers, S. Laurence & S. Stich (Eds.), *The innate mind: Structure and content* Oxford: Oxford University Press., pp. 53-68.

Sperber, D., & Hirschfeld, L.A. (2004). The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences, 8*(1), pp. 40-46.

Spreng, R.N., Mar, R.A., & Kim, A.S.N. (2009). The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), pp. 489-510.

Stacy, A.W., & Wiers, R.W. (2010). Implicit Cognition and Addiction: A Tool for Explaining Paradoxical Behavior. *Annual Review of Clinical Psychology, 6*(1), pp. 551-575.

Stanovich, K.E. (2005). *The robot's rebellion: finding meaning in the age of Darwin*. Chicago: University Of Chicago Press.

Stanovich, K.E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J.S.B.T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond.* Oxford: Oxford University Press, pp. 55-88.

Steele, C.J., & Penhune, V.B. (2010). Specific Increases within Global Decreases: A Functional Magnetic Resonance Imaging Investigation of Five Days of Motor Sequence Learning. *J. Neurosci., 30*(24), pp. 8332-8341.

Stenning, K., & Lambalgen, M.v. (2008). *Human reasoning and cognitive science.* Cambridge, MA: MIT Press.

Stephen, D.G., Dixon, J.A., & Isenhower, R.W. (2009). Dynamics of Representational Change:

Entropy, Action, and Cognition. *Journal of Experimental Psychology-Human Perception and Performance, 35*(6), pp. 1811-1832.

Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), pp. 465-481.

Stokhof, M. (2000). *Taal en betekenis. Een inleiding in de taalfilosofie*. Amsterdam: Boom.

Stout, D. (2010). The Evolution of Cognitive Control. *Topics in Cognitive Science, 2*(4), pp. 614-630.

Stout, D., & Chaminade, T. (2009). Making Tools and Making Sense: Complex, Intentional Behaviour in Human Evolution. *Cambridge Archaeological Journal, 19*(01), pp. 85-96.

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review, 8*(3), pp. 220-247.

Straube, B. (2012). An overview of the neuro-cognitive processes involved in the encoding, consolidation, and retrieval of true and false memories. *Behavioral and Brain Functions, 8*.

Strawson, G. (2004). Against Narrativity. *Ratio, 17*(4), pp. 428-452.

Suddendorf, T. (2010). Episodic memory versus episodic foresight: similarities and differences. *Wiley Interdisciplinary Reviews-Cognitive Science, 1*(1), pp. 99-107.

Suddendorf, T., & Corballis, M.C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences, 30*(03), pp. 299-313.

Sullivan, J.A. (2010). Reconsidering 'spatial memory' and the Morris water maze. *Synthese, 177*(2), pp. 261-283.

Sumner, P., & Husain, M. (2008). At the Edge of Consciousness: Automatic Motor Activation and Voluntary Control. *Neuroscientist, 14*(5), pp. 474-486.

Sun, R., Coward, L.A., & Zenzen, M.J. (2005). On levels of cognitive modeling. *Philosophical psychology, 18*(5), pp. 613-637.

Sun, R., Slusarz, P., & Terry, C. (2005). The Interaction of the Explicit and the Implicit in Skill Learning: A Dual-Process Approach. *Psychological Review, 112*(1), pp. 159-192.

Tanji, J., & Hoshi, E. (2001). Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology, 11*(2), pp. 164-170.

Taubert, M., Lohmann, G., Margulies, D.S., et al. (2011). Long-term effects of motor training on resting-state networks and underlying brain structure. *NeuroImage, 57*(4), pp. 1492-1498.

Taylor, L.J., & Zwaan, R.A. (2009). Action in cognition: The case of language. *Language and Cognition, 1*(1), pp. 45-58.

Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science, 22*(1), pp. 1-24.

Thakkar, K.N., Nichols, H.S., McIntosh, L.G., et al. (2011). Disturbances in Body Ownership in Schizophrenia: Evidence from the Rubber Hand Illusion and Case Study of a Spontaneous

Out-of-Body Experience. *PLoS ONE, 6*(10), pp. 1-10.

Thomas, M., & Karmiloff-Smith, A. (1998). Quo vadis modularity in the 1990S? *Learning and Individual Differences, 10*(3), pp. 245-250.

Thompson, E. (1995). *Colour vision: A study in cognitive science and the philosophy of perception*. London: Routledge.

Thompson, E., & Stapleton, M. (2009). Making Sense of Sense-Making: Reflections on Enactive and Extended Mind Theories. *Topoi, 28*(1), pp. 23-30.

Tilly, C. (2001). Mechanisms in political processes. *Annual Review of Political Science, 4*(1), pp. 21-41.

Tindell, A.J., Smith, K.S., Berridge, K.C., et al. (2009). Dynamic Computation of Incentive Salience:" Wanting" What Was Never" Liked". *Journal of Neuroscience, 29*(39), pp. 12220.

Tononi, G., & Koch, C. (2008). The Neural Correlates of Consciousness: An Update. *Annals of the New York Academy of Sciences, 1124*(1), pp. 239.

Trabasso, T., & Stein, N.L. (1994). Using goal-plan knowledge to merge the past with the present and the future in narrating events on line. In M. Maith, J.B. Benson, R.J. Roberts & B.F. Pennington (Eds.), *The development of future oriented processes* Chicago: University of Chicago Press, pp. 323-349.

Tracy, J., Flanders, A., Madi, S., et al. (2003). Regional Brain Activation Associated with Different Performance Patterns during Learning of a Complex Motor Skill. *Cerebral Cortex, 13*(9), pp. 904-910.

Travassos, B., Araújo, D., Davids, K., et al. (2013). Expertise effects on decision-making in sport are constrained by requisite response behaviours – A meta-analysis. *Psychology of Sport and Exercise, 14*(2), pp. 211-219.

Trehub, S.E., & Hannon, E.E. (2006). Infant music perception: Domain-general or domain-specific mechanisms? *Cognition, 100*(1), pp. 73-99.

Trehub, S.E., Schellenberg, E.G., & Nakata, T. (2008). Cross-cultural perspectives on pitch memory. *Journal of Experimental Child Psychology, 100*(1), pp. 40-52.

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1373), pp. 1295-1306.

Tremblay, P., & Small, S.L. (2011). From Language Comprehension to Action Understanding and Back Again. *Cerebral Cortex, 21*(5), pp. 1166-1177.

Turnbull, W., & Carpendale, J.I.M. (2009). Talk and Childrens Understanding of Mind. *Journal of Consciousness Studies, 16*(6-8), pp. 140-166.

Tylen, K., Weed, E., Wallentin, M., et al. (2010). Language as a Tool for Interacting Minds. *Mind & Language, 25*(1), pp. 3-29.

Tzeng, Y., Broek, P., Kendeou, P., et al. (2005). The computational implementation of the

landscape model: Modeling inferential processes and memory representations of text comprehension. *Behavior Research Methods, 37*(2), pp. 277-286.

Uithol, S., van Rooij, I., Bekkering, H., et al. (2012). Hierarchies in Action and Motor Control. *Journal of Cognitive Neuroscience, 24*(5), pp. 1077-1086.

Umiltà, M.A., Escola, L., Intskirveli, I., et al. (2008). When pliers become fingers in the monkey motor system. *Proceedings of the National Academy of Sciences, 105*(6), pp. 2209-2213.

Uttal, W.R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain.* Cambridge, MA: MIT Press.

Valet, M., Sprenger, T., Boecker, H., et al. (2004). Distraction modulates connectivity of the cingulo-frontal cortex and the midbrain during pain: an fMRI analysis. *Pain, 109*(3), pp. 10.

Valyear, K.F., Gallivan, J.P., McLean, D.A., et al. (2012). fMRI Repetition Suppression for Familiar But Not Arbitrary Actions with Tools. *The Journal of Neuroscience, 32*(12), pp. 4247-4259.

van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: The role of co-activation in confronting misconceptions. *Applied Cognitive Psychology, 22*(3), pp. 335-351.

van der Eijk, P.J. (1997). The Matter of Mind: Aristotle on the Biology of Psychic Processes. In W. Kullmann & S. Follinger (Eds.), *Aristotelische Biologie. Intentionen, Methoden, Ergebnisse.* Stuttgart: Steiner Verlag, pp. 231-241.

van der Lecq, R. (2012). Why we talk: an interdisciplinary approach to the evolutionary origin of language. In A. Repko, W.H. Newell & R. Szostak (Eds.), *Interdisciplinary research: Case studies of interdisciplinary understandings of complex problems* Thousand Oaks, CA: SAGE, pp. 191-223.

Van Dijk, T.A. (1976). Philosophy of action and theory of narrative. *Poetics, 5*(4), pp. 287-338.

van Gaal, S., & Lamme, V.A.F. (2012). Unconscious High-Level Information Processing: Implication for Neurobiological Theories of Consciousness. *The Neuroscientist, 18*(3), pp. 287-301.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21*(05), pp. 615-628.

Van Gelder, T., & Port, R.F. (1995). It's about time: An overview of the dynamical approach to cognition. In R.F. Port & T. Van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* Cambridge, MA: MIT Press, pp. 1-43.

van Mier, H., Tempel, L.W., Perlmutter, J.S., et al. (1998). Changes in Brain Activity During Motor Learning Measured With PET: Effects of Hand of Performance and Practice. *J Neurophysiol, 80*(4), pp. 2177-2199.

Vanderwolf, C.H. (1998). Brain, behavior, and mind: What do we know and what can we

know? *Neuroscience and biobehavioral reviews 22*(2), pp. 125-142.

Velleman, J.D. (2003). Narrative Explanation. *The Philosophical Review, 112*(1), pp. 1-25.

Verano, J.W., & Finger, S. (2009). Ancient trepanation. In M.J. Aminoff, F. Boller & D.F. Swaab (Eds.), *Handbook of Clinical Neurology* Vol. Volume 95: Elsevier, pp. 3-14.

Verwey, W. (2003). Processing modes and parallel processors in producing familiar keying sequences. *Psychological Research, 67*(2), pp. 106-122.

Vicente, K.J., & Wang, J.H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105*(1), pp. 33-57.

Vogeley, K., & Roepstorff, A. (2009). Contextualising culture and social cognition. *Trends in Cognitive Sciences, 13*(12), pp. 511-516.

Vogt, S., Buccino, G., Wohlschla?ger, A.M., et al. (2007). Prefrontal involvement in imitation learning of hand actions: Effects of practice and expertise. *NeuroImage, 37*(4), pp. 1371.

Walter, H. (2001). *Neurophilosophy of free will. From libertarian illusions to a concept of natural autonomy* (C. Klohr, Trans.). Cambridge, MA: MIT Press.

Wason, P.C., & Evans, J.S.B. (1975). Dual processes in reasoning? *Cognition, 3*(1), pp. 41-54.

Webb, T.L., & Sheeran, P. (2007). How do implementation intentions promote goal attainment? A test of component processes. *Journal of Experimental Social Psychology, 43*(2), pp. 295-302.

Weber, A., & Vosgerau, G. (2012). Grounding Action Representations. *Review of Philosophy and Psychology, 3*(1), pp. 53-69.

Weigelt, M., Ahlmeyer, T., Lex, H., et al. (2011). The cognitive representation of a throwing technique in judo experts – Technological ways for individual skill diagnostics in high-performance sports. *Psychology of Sport and Exercise, 12*(3), pp. 231-235.

Weiskopf, D.A. (2010). The Goldilocks problem and extended cognition. *Cognitive Systems Research, 11*(4), pp. 313-323.

Weiskrantz, L. (1997). *Consciousness lost and found: a neuropsychological exploration*. Oxford: Oxford University Press.

West, M.L. (1992). *Ancient greek music*: Oxford University Press.

Westermann, G., Mareschal, D., Johnson, M.H., et al. (2007). Neuroconstructivism. *Developmental Science, 10*(1), pp. 75-83.

Wheeler, M. (2010). The Problem of Representation. In D. Schmicking & S. Gallagher (Eds.) *Handbook of Phenomenology and Cognitive Science*. Dordrecht: Springer, pp. 318-336.

Wheeler, M., & Clark, A. (2008). Culture, embodiment and genes: unravelling the triple helix. *Philos. Trans. of the Royal Society B: Biological Sciences, 363*(1509), pp. 3563-3575.

White, R.C., Davies, A.M.A., Halleen, T.J., et al. (2010). Tactile expectations and the perception of self-touch: An investigation using the rubber hand paradigm. *Consciousness and Cognition, 19*(2), pp. 505-519.

Whitley, D.S. (1998). Cognitive neuroscience, shamanism and the rock art of native California. *Anthropology of Consciousness, 9*(1), pp. 22-37.

Willems, R.M. (2009). *Neural reflections of meaning in gesture, language, and action*, Radboud Universiteit Nijmegen. Retrieved on August 30, 2013 from http://webdoc.ubn.ru.nl/ mono/w/willems_r/neurreofm.pdf

Willems, R.M., & Hagoort, P. (2007). Neural evidence for the interplay between language, gesture, and action A review. *Brain and language, 101*(3), pp. 278-289.

Williamon, A., & Valentine, E. (2002). The Role of Retrieval Structures in Memorizing Music. *Cognitive Psychology, 44*(1), pp. 1-32.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), pp. 625-636.

Wilson, R.A. (1994). Wide Computationalism. *Mind, 103*(411), pp. 351-372.

Wilson, S.M., Molnar-Szakacs, I., & Iacoboni, M. (2008). Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension. *Cereb. Cortex, 18*(1), pp. 230-242.

Wimsatt, W.C. (1976). Reductionism, levels of organization, and the mind-body problem. In G. Globus, G. Maxwell & I. Savodnik (Eds.), *Consciousness and the brain: A scientific and philosophical inquiry* New York: Plenum, pp. 205-267.

Wimsatt, W.C. (1986). Developmental Constraints, Generative Entrenchment, and the Innate-Acquired Distinction. In W. Bechtel (Ed.), *Integrating Scientific Disciplines*, Dordrecht: Martinus Nijhoff, pp. 185-208.

Wimsatt, W.C. (1999). Genes, Memes, and Cultural Heredity. *Biology and Philosophy, 14*, pp. 279-310.

Wimsatt, W.C. (2001). Generative entrenchment and the developmental systems approach to evolutionary processes. In S. Oyama, P.E. Griffiths & R.D. Gray (Eds.), *Cycles of Contingency: Developmental Systems and Evolution* Cambridge, MA: MIT Press, pp. 219-237.

Wimsatt, W.C. (2006). Aggregate, composed, and evolved systems: Reductionistic heuristics as means to more holistic theories. *Biology and Philosophy, 21*(5), pp. 667.

Wimsatt, W.C. (2006). Generative entrenchment and an evolutionary developmental biology for culture. *Behavioral and Brain Sciences, 29*(04), pp. 364-366.

Wimsatt, W.C. (2007). *Re-Engineering Philosophy for Limited Beings. Piecewise approximations to reality*: Harvard University Press.

Wimsatt, W.C., & Griesemer, J.R. (2007). Reproducing Entrenchments to Scaffold Culture: The Central Role of Development in Cultural Evolution. In R. Sansom & R.M. Brandon (Eds.), *Integrating Evolution and Development: From Theory to Practice*. Cambridge, MA: MIT Press, pp. 227-323.

Wimsatt, W.C., & Schank, J.C. (2004). Generative entrenchment, modularity, and evolvability: When genic selection meets the whole organism. In G. Schlosser & G.P. Wagner (Eds.), *Modularity in development and evolution* Chicago: Chicago University Press, pp. 359-394.

Winkelman, M. (2003). Evolutionary and Neurohermeneutic Approaches to Culture and the Brain. *Reviews in Anthropology, 32*(4), pp. 275-295.

Witt, J.K., Proffitt, D.R., & Epstein, W. (2005). Tool use affects perceived distance, but only when you intend to use it. *Journal of Experimental Psychology-Human Perception and Performance, 31*(5), pp. 880-888.

Woerther, F. (2008). Music and the education of the soul in Plato and Aristotle: Homeopathy and the formation of character. *The Classical Quarterly, 58*(01), pp. 89-103.

Wolpert, D., Ghahramani, Z., & Jordan, M. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), pp. 1880-1882.

Wolpert, D.M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience, 3*(11), pp. 1212-1217.

Wood, J.N., Tierney, M., Bidwell, L.A., et al. (2005). Neural correlates of script event knowledge: A neuropsychological study following prefrontal injury. *Cortex, 41*(6), pp. 796-804.

Wood, W., & Neal, D.T. (2007). A New Look at Habits and the Habit–Goal Interface. *Psychological Review, 114*(4), pp. 843-863.

Woodward, J. (2002). What Is a Mechanism? A Counterfactual Account. *Philosophy of Science, 69*(3), pp. S366-S377.

Xu, F. (2002). Language and conceptual development: Words as essence placeholders. *Behavioral and Brain Sciences, 25*(06), pp. 704-705.

Xu, J., Kemeny, S., Park, G., et al. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage, 25*(3), pp. 1002-1015.

Xu, T., Yu, X., Perlik, A.J., et al. (2009). Rapid formation and selective stabilization of synapses for enduring motor memories. *Nature, 462*, pp. 415-419.

Yarkoni, T., Speer, N.K., & Zacks, J.M. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage, 41*(4), pp. 1408-1425.

Yin, H.H., & Knowlton, B.J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7*(6), pp. 464-476.

Yin, H.H., Mulcare, S.P., Hilário, M.R., et al. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nat Neurosci, 12*, pp. 333-341.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), pp. 301-308.

Zacks, J.M., Kumar, S., Abrams, R.A., et al. (2009). Using movement and intentions to understand human activity. *Cognition, 112*(2), pp. 201-216.

Zacks, J.M., & Sargent, J.Q. (2010). Event Perception: A Theory and Its Application to Clinical

Neuroscience. In H.R. Brian (Ed.), *Psychology of Learning and Motivation* Vol. Volume 53: Academic Press, pp. 253-299.

Zacks, J.M., Speer, N.K., Swallow, K.M., et al. (2007). Event Perception: A Mind-Brain Perspective. *Psychological Bulletin, 133*(2), pp. 273-293.

Zanini, S., Rumiati, R.I., & Shallice, T. (2002). Action Sequencing Deficit Following Frontal Lobe Lesion. *Neurocase, 8*(1-2), pp. 88-99.

Zelazo, P.D., & Frye, D. (1998). Cognitive Complexity and Control: II. The Development of Executive Function in Childhood. *Current Directions in Psychol. Science, 7*(4), pp. 121-126.

Zeman, A. (2001). Consciousness. *Brain, 124*(7), pp. 1263-1289.

Zhu, Y., Zhang, L., Fan, J., et al. (2007). Neural basis of cultural influence on self-representation. *NeuroImage, 34*(3), pp. 1310-1316.

Zurbriggen, E.L., Fontenot, D.L., & Meyer, D.E. (2006). Representation and execution of vocal motor programs for expert singing of tonal melodies. *Journal of Experimental Psychology: Human Perception and Performance, 32*(4), pp. 944-963.

Zwaan, R.A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *The psychology of learning and motivation, 44*, pp. 35-62.

Zwaan, R.A. (2008). Time in Language, Situation Models, and Mental Simulations. *Language Learning, 58*(s1), pp. 13-26.

Zwaan, R.A. (2009). Mental simulation in language comprehension and social cognition. *European Journal of Social Psychology, 39*(7), pp. 1142-1150.

Zwaan, R.A., Langston, M.C., & Graesser, A.C. (1995). The construction of situation models in narrative comprehension: An Event-Indexing Model. *Psychol. Science, 6*(5), pp. 292-297.

Zwaan, R.A., Stanfield, R.A., & Madden, C.J. (1999). Perceptual symbols in language comprehension: Can an empirical case be made? *Behavioral and Brain Sciences, 22*(04), pp. 636-637.

## SAMENVATTING. FILOSOFISCHE EN COGNITIEF NEUROWETENSCHAPPELIJKE INZICHTEN IN HET VORMEN VAN DE 'HANDELINGS-RUIMTE' DOOR EEN EXPERT*

Waarom loven wij personen die bepaalde handelingen op uiterst bekwame wijze uitvoeren zonder daarover te hoeven nadenken, terwijl wij onze lof normaliter voorbehouden aan weloverwogen en bewust gecontroleerde handelingen? Men kan daarom wel spreken van een 'paradox van bekwaam handelen': men zou denken dat beginnelingen onze lof verdienen omdat zij tenminste elk handelings-deel moeten bedenken en controleren, ook al handelen zij zeer matig en inflexibel. Dat zou echter wel impliceren dat wij personen steeds minder zouden prijzen naarmate zij hun handelingen minder bedachtzaam en bewust uitvoeren en juist meer op de automatische piloot. De professionele operazanger zou dan juist minder geprezen worden dan het kind dat zijn zangstem nog moet ontdekken.

Deze paradox van bekwaam handelen heeft nog andere facetten. Niet alleen wordt dezelfde handeling door beginnelingen en door experts op verschillende manieren uitgevoerd, met zichtbaar verschillend resultaten. Experts zijn ook beter in staat om hun handelingen gericht en minutieus te verbeteren en over te dragen aan een leerling. Cognitie(neuro-)-wetenschappelijk onderzoek toont bovendien aan dat de hersenprocessen die betrokken zijn bij dergelijke handelingen in veel opzichten verschillen tussen beginnelingen en experts. Ontwikkeling en leren hebben blijkbaar ingrijpende gevolgen voor zowel de uiterlijke kenmerken van een handeling als voor de onderliggende hersenprocessen.

Om deze situatie en paradox te analyseren moeten we navigeren tussen conceptuele en empirische inzichten, waarbij recht gedaan moet worden aan zowel de overeenkomsten tussen de handelingen van beginnelingen en experts als aan de verschillen daartussen. In dit proefschrift (zie § III.1.1) wordt daartoe een concept geïntroduceerd dat helpt bij de analyse en verklaring van die overeenkomsten en verschillen, namelijk het begrip van een 'handelingsruimte' die 'geboetseerd' of gevormd kan worden (een zgn. 'sculpted space of actions'). Daarbij stellen we ons voor dat wanneer iemand in een bepaalde situatie moet gaan handelen, die persoon meestal zou kunnen kiezen uit meerdere mogelijke handelingen die hij kan uitvoeren. Dan kunnen we elke handelingsoptie beschouwen als een afzonderlijk punt of gebiedje in een

---

* Een uitvoerige Engelstalige samenvatting is hierboven te vinden onder de titel: "Conclusion and summary. Why sculpting the space of action matters." Verder wordt de lezer verwezen naar drie figuren die na die samenvatting op de pagina's 371-375 zijn opgenomen en die vereenvoudigde representaties bieden van de inhoud van de drie delen van dit proefschrift.

persoonlijke, meer-dimensionale handelingsruimte waaruit die persoon moet putten. De dimensies van die handelingsruimte zijn natuurlijk niet de dimensies van hoogte, breedte en lengte, maar veelmeer dimensies die bijvoorbeeld zouden kunnen weergeven of een handelingsoptie adequaat is in die bepaalde situatie, of die optie makkelijk uitvoerbaar is door die persoon en of die optie ook past bij diens overige plannen en intenties.

Conform dit concept van een meer-dimensionele handelingsruimte kunnen we ons ook voorstellen dat de ene optie sneller geselecteerd en uitgevoerd zal worden dan de andere, afhankelijke van de plaats en omvang ervan in iemands handelingsruimte. Een veel geoefende en geprefereerde handeling zal bijvoorbeeld een prominentere en grotere plaats in die handelingsruimte innemen dan een ongewone en verafschuwde handeling, die slechts perifeer en minimaal in die handelingsruimte gerepresenteerd wordt en dus niet snel geselecteerd en uitgevoerd zal worden. Iemands handelingsruimte is natuurlijk afhankelijk van lange-termijn processen zoals zijn ontwikkeling en zijn lange-termijn intenties, maar ook van factoren die te maken hebben met de concrete situatie waarin hij zich bevindt. Wanneer iemand expertise vergaart dan is hij – of zij – welbeschouwd bezig met het boetseren van zijn handelingsruimte waardoor hij steeds weer zo snel, flexibel en adequaat handelt. De handelingsruimte van een geoefende operazanger die Don Giovanni zingt zal er dus heel anders uitzien dan de beginneling die onvoorbereid op het podium geduwd wordt en voor wie er dan nauwelijks handelingsopties beschikbaar zijn die muzikaal en theatraal adequaat zijn. Eventueel zou de beginneling de juiste noten kunnen zingen en gebaren maken, maar dat zou eerder het resultaat van stom toeval en geluk zijn dan van hun prominente plaats in zijn handelingsruimte op dat moment.

Dit concept van de 'geboetseerde handelingsruimte' wordt ontwikkeld in de drie delen waaruit dit proefschrift bestaat: een methodologisch deel over cognitiewetenschappelijke verklaringen, een deel waarin theorieën over ontwikkeling en leerprocessen centraal staan en tenslotte een deel waarin filosofische en cognitiewetenschappelijke inzichten omtrent menselijk handelen worden geïntegreerd. De vier cognitiewetenschappelijke verklaringsmethodes die in deel I worden besproken geven elk een eigen visie op de wijze waarop onderzoekers concepten, definities en empirische evidentie met elkaar moeten verbinden. Willen wetenschappers bijvoorbeeld inzicht verwerven in de processen die betrokken zijn bij bewustzijn, emoties, waarneming of gedrag, dan moeten zij het wel met elkaar eens zijn wanneer er sprake van zo'n functie is en wanneer niet. Kan bewustzijn worden onderzocht bij dieren, of bij patiënten in een vegetatieve toestand, bijvoorbeeld, of beschikken die daar niet over? Conceptuele onenigheid of onduidelijkheid kan verstrekkende gevolgen hebben omdat dan

onzeker is of onderzoekers wel echt hetzelfde fenomeen of proces onderzoeken, ook al geven ze daaraan dezelfde naam. Onderzoeksresultaten kunnen daarmee hun relevantie verliezen.

Begonnen wordt in hoofdstuk I.2 met een kritische bespreking van de opvatting van Bennett & Hacker, volgens welke een conceptuele analyse van een cognitieve functie tot eenduidige en richtinggevende resultaten voor empirisch onderzoek zou moeten leiden. Deze veronderstelling blijkt echter niet houdbaar en in plaats daarvan wordt voorgesteld om conceptuele inconsistenties of verrassende en afwijkende begrippen juist te gebruiken als heuristiek, als inspiratie voor vervolgonderzoek. De methode die voorgesteld werd door David Marr en besproken wordt in I.3 sluit beter bij dit laatste voorstel aan. Hij betoogde immers dat de analyse en verklaring van een cognitieve functie aandacht moet besteden aan drie heel verschillende perspectieven daarop: een taakanalyse van die functie, een analyse van de manieren waarop de informatie die daarbij betrokken is kan worden weergegeven en suggesties voor de hersenprocessen die dan uiteindelijk de functie realiseren. Anders dan Marr bepleitte wordt door ons betoogd dat de integratie van deze perspectieven juist tot een plausibele en robuuste verklaring van een functie kan leiden. Onderzoekers die op zoek zijn naar neurale correlaten van bewustzijn (Neural Correlates of Consciousness) vereisen niet zo'n integratie maar laten een veel lossere relatie toe tussen empirisch onderzoek en eventuele definities of conceptuele analyses van een functie. In hoofdstuk I.4 wordt betoogd dat voor de zoektocht naar correlaties tussen een bepaalde functie en geobserveerde hersenprocessen er toch criteria nodig zijn om te besluiten of een bepaalde vondst werkelijk slaat op de functie die men probeert te onderzoeken. Deze methode kan dus vruchtbaar zijn, maar slechts als een eerste stap bij het ontwikkelen van een meer omvattende verklaring.

Na deze drie verklaringsmethoden wordt in I.5 de zogenaamde 'mechanistische verklaringsmethode' geïntroduceerd en uiteengezet. Deze methode blijkt uitstekend geschikt om de resultaten van heel verschillende typen van onderzoek te integreren en lijkt daarmee Marr's methode verder uit te werken. Daartoe worden een aantal heuristieken of strategieën gehanteerd, die een taakverdeling tussen wetenschappers mogelijk maakt en die hen ook in staat stelt om eerdere resultaten bij te stellen in het licht van nieuwere inzichten. Om te beginnen moet het te verklaren fenomeen, een cognitieve functie bijvoorbeeld, nader bepaald worden om verwarring met andere functies zoveel mogelijk te voorkomen. Vervolgens wordt zo'n functie opgedeeld in deel-taken (of deel-functies) die deels apart kunnen worden onderzocht, zoals bijvoorbeeld waarneming, geheugen, stembeheersing enzovoorts in het geval van onze Don Giovanni. Natuurlijk bestaan er interacties tussen die deel-taken

maar in eerste instantie zullen onderzoekers proberen die afzonderlijk ergens in de hersens te lokaliseren. Wanneer veel meer inzicht verkregen is in het complexe verklaringsmechanisme van een functie, dan kunnen onderzoekers overwegen om de definitie van die functie enigszins daaraan aan te passen. Niet alleen is het opstellen van een verklarend mechanisme van een functie vruchtbaar voor de integratie van filosofische en cognitiewetenschappelijke inzichten, de veranderingen onder invloed van ontwikkeling en leren kunnen zo ook worden verklaard.

Na deze methodologische beschouwing in deel I is deel II gewijd aan de bespreking van een aantal cognitief neurowetenschappelijke theorieën over ontwikkeling en leren. Daarbij worden ook de verklaringsmechanismen onderzocht die bij deze theorieën horen, om zo inzicht te krijgen in hetgeen er plaatsvindt wanneer iemand expertise vergaart. Meer in het algemeen wordt in navolging van Herbert Simon en Wimsatt gesteld dat het voor dynamische systemen voordelig is wanneer zij in staat zijn tot structurele en stabiele veranderingen omdat zij daardoor efficiënter en adequater kunnen opereren en vervolgens steeds complexere taken kunnen leren en uitvoeren. Alleen met een geoefende stem heeft het zin om een opera-rol te gaan spelen, omdat alleen dan de benodigde aandacht voor het theatrale spel opgebracht kan worden. Het effect van deze ontwikkeling is vaak dat er in het verantwoordelijke mechanisme een zogenaamde 'kludge' gevormd wordt: sommige componenten van zo'n mechanisme vertonen dan meer interactie met elkaar terwijl andere componenten overbodig geworden zijn. De structuur van hersenprocessen verandert bijvoorbeeld onder invloed van ontwikkeling en leren, zodat een expert een functie kan uitoefenen met veel minder en minder verspreide hersenactiviteit. In dit deel wordt ook onderzocht of dergelijke structurele veranderingen aan bod komen in de besproken theorieën.

In hoofdstuk II.2 wordt ingegaan op de 'neuroconstructivistische' theorie van ontwikkelen en leren, onder andere van Karmiloff-Smith. Volgens deze theorie zijn er twee processen die in dat verband van belang zijn. Ten eerste wordt een bepaalde taak steeds meer beheerst als een relatief eenvoudige procedure, die weinig aandacht behoeft. Ten tweede verkrijgt een leerling steeds meer expliciete controle over zo'n taak, zodat hij gericht zijn resultaten kan beïnvloeden. Volgens het neuroconstructivisme is dit alles in belangrijke mate het gevolg van het feit dat het verklarende mechanisme een complexere en meer modulaire structuur krijgt, daarmee onze verwachting bevestigend.

Daarna wordt in II.3 aandacht besteed aan zogenaamde 'duale-proces theorieën'. Deze theorieën komen overeen in het feit dat zij onderscheid maken tussen een

automatische en een meer gecontroleerde manier van het uitvoeren van een bepaalde functie of taak, ondersteund door verschillende hersenprocessen. Tussen deze twee uitvoeringswijzen bestaan allerlei verschillen en de vraag is dan ook of een expert enigszins kan beheersen op welke manier hij een taak uitvoert. Wij wijzen erop dat zulke zelfregulering in enige mate zeker mogelijk is en dat zelfregulering soms ook weer geautomatiseerd kan worden, zodat een expert kan voorkomen dat hij steeds in bepaalde situaties een ongewenste, automatische handeling vertoont.

Omdat uiteindelijk in dit proefschrift ook de rol van bewuste en talige intenties onderzocht wordt staan we in II.4 stil bij Barsalou's simulatie-theorie. De informatie die behoort bij een bepaalde handeling of een bepaald begrip, bijvoorbeeld, wordt in vele onderdelen en verspreid over het brein opgeslagen: de beelden die iemand heeft van Don Giovannie worden op een andere plaats opgeslagen dan de herinnerde melodieën en die weer op andere plaatsen dan de motorpatronen die passen bij het machismo van deze verleider. De geboetseerde handelingsruimte van een expert wordt deels bepaald door de 'simulatoren' die hij in de loop der tijd heeft ontwikkeld en die steeds makkelijk geactiveerd kunnen worden. Hierbij aansluitend wordt kort de 'extended cognition' theorie besproken, die nog nadrukkelijker toont hoezeer externe informatie – taal, gereedschap – geïntegreerd kan raken in dit soort simulatoren.

Al met al bevestigt deel II dat ontwikkeling en leren begrepen en verklaard kunnen worden als veranderingen die een verklarend mechanisme ondergaat. Gerichte intenties van een expert kunnen die stabiele veranderingen op verschillende manieren beïnvloeden. Bovendien is het niet zo dat een expert alle controle verliest over de manier waarop zijn handeling uitgevoerd wordt. In deel III zullen we nader onderzoeken of de paradox van bekwaam handelen inderdaad beperkte geldigheid heeft omdat expertise onder meer bestaat uit het boetseren van een handelingsruimte die tot dan toe vrijwel ongevormd was.

Deel III richt zich op de opzettelijke of intentionele handeling, waarbij gebruik wordt gemaakt van de methodologische inzichten uit deel I en de inzichten met betrekking tot de ontwikkeling en het leren uit deel II. Intentioneel handelen kan namelijk in allerlei opzichten vergeleken worden met bekwaam handelen of expertise. In beide gevallen is namelijk het 'boetseren van de handelingsruimte' van belang, hetgeen in III.1 wordt uitgelegd. De verdere uiteenzetting maakt gebruik van Pacherie's theoretisch kader omtrent intentioneel handelen, dat zelf weer berust op filosofische en cognitiewetenschappelijke inzichten in dergelijk handelen. Dit kader is in belangrijke mate beïnvloed door de filosofen Frankfurt en Bratman en de cognitiewetenschapper Jeannerod, van wie het werk ook in dit hoofdstuk besproken

wordt – naast dat van vele anderen.

Pacherie's kader van de 'intentionele cascade' onderscheidt drie niveaus van intenties, die op verschillende manieren met elkaar interacteren en zo uiteindelijk een handeling mogelijk maken. Onderaan de hiërarchie staan de motor intenties, waarmee de representaties bedoeld worden die ten grondslag liggen aan de motorische bewegingen waarmee een handeling uitgevoerd wordt. Een motor intentie moet natuurlijk passen bij zowel de concrete situatie en de objecten die daarin voorhanden zijn, als ook bij iemands distale intenties of lange-termijn plannen, die bovenaan de hiërarchie staan. In het midden bevinden zich de proximale of korte-termijn intenties die dan ook verantwoordelijk zijn voor het verankeren van een distale intentie in een concrete situatie en de specificatie daarvan zodanig dat een motor intentie gevormd kan worden. Conform deze drie niveaus van intenties wordt deel III opgebouwd, waarbij steeds per niveau eerst een filosofische analyse geboden wordt en vervolgens onderzocht in hoeverre cognitiewetenschappelijke inzichten deze analyse ondersteunen, of dat wederzijdse bijstelling nodig is.

Een belangrijk resultaat van de filosofische analyses in dit gehele deel is dat niet alleen om morele redenen het van belang is dat een persoon een relatief stabiele hiërarchie van intenties en voorkeuren ontwikkelt. Frankfurt en Bratman hebben verschillende argumenten hiervoor ontwikkeld. Wanneer een persoon steeds weer, als ware het voor het eerst, een besluit moet nemen om een bepaalde handeling al dan niet uit te voeren, dan zal hij aan dergelijke besluiten steeds veel tijd en aandacht moeten besteden. Bovendien maken de meeste handelingen deel uit van allerlei samengestelde handelingen en vergen een langere periode voor hun uitvoering. Heroverweging van een lange-termijn intentie zou betekenen dat vele eerder uitgevoerde handelingen zinloos worden en misschien zelfs contra-productief. De betekenis van een handeling en de inbedding ervan in een groter geheel hangen dus nauw samen. Aan de hand van Ricoeurs analyses wordt tenslotte gewezen op het belang van de 'narratieve simulatie' van handelingen, omdat daarmee een dergelijke inbedding ontwikkeld kan worden. Onze operazanger zou bijvoorbeeld het podium nooit hebben bereikt als hij elk uur van de dag weer opnieuw had overwogen of hij deze toonladders wel echt wilde studeren. Zijn uiteindelijke intentie kon dus alleen gerealiseerd worden doordat hij allerlei daaraan ondergeschikte intenties heeft opgenomen in een geboetseerde handelingsruimte.

Uit de veelsoortige cognitiewetenschappelijke evidentie die in dit deel besproken wordt blijkt eveneens hoe belangrijk het is dat ons brein – en ons lichaam in meer algemene zin – in staat is om nieuwe en stabiele onderdelen te ontwikkelen die complexe handelingen mogelijk maken. Zo blijkt het brein complexe informatie op

verschillende manieren te kunnen comprimeren, als de onderdelen van die informatie maar vaak genoeg en tegelijkertijd wordt verwerkt. Dit fenomeen is terug te zien bij het groeperen of 'chunken' van informatie, waardoor experts in een oogopslag zeer complexe informatie kunnen doorgronden en daarop adequaat kunnen reageren. Iets soortgelijks geldt ook voor de ontwikkeling van handelingsschema's, die samengesteld zijn uit allerlei deelhandelingen maar toch als een geheel vlot en zonder moeite geactiveerd kunnen worden. Zo kan de handeling behorend bij een bepaalde aria geactiveerd worden in het geval dat onze operazanger slechts de inzet ervan hoort, maar hij kan die ook actief oproepen door zich de tekst van die aria te herinneren.

Al navigerend tussen filosofische analyse en empirische evidentie wordt in dit deel betoogd dat menselijk, intentioneel handelen in allerlei opzichten te vergelijken is met het vergaren van expertise. Door een langdurig en deels weloverwogen proces stelt een expert zichzelf in staat om intentionele handelingen uit te voeren die complexer en adequater zijn dan die van een beginneling. Dankzij een geboetseerde handelingsruimte kan een persoon dus niet alleen operazanger worden maar kan hij ook op adequate wijze morele handelingen uitvoeren, mits de bijbehorende morele intenties in die handelingsruimte zijn geïntegreerd. De drie delen van dit proefschrift hebben betoogd dat voor de verklaring van dit opmerkelijke verschijnsel gewerkt moet worden aan de integratie van verklarende mechanismes en intentionele structuren, terwijl daarbij ook rekening gehouden moet worden met ontwikkeling en leren. Gezien deze complexiteit lijkt ons de bewondering die gekoesterd wordt voor een operazanger of een morele held zeker gerechtvaardigd en zij zouden ons moeten inspireren om onze eigen handelingsruimte te gaan boetseren.