Tales of
Similarity and Imagination

Tom Schoonen

Tales of Similarity and Imagination

A modest epistemology of possibility

Tom Schoonen

Tales of Similarity and Imagination

A modest epistemology of possibility

ILLC Dissertation Series DS-2020-15

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Thom van Gessel & Tom Schoonen
Printed and bound by Ipskamp Printing.

ISBN: 978-94-6421-096-5

Tales of Similarity and Imagination

—

A modest epistemology of possibility

Academisch Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 4 december 2020, te 16.00 uur

door

Tom Schoonen

geboren te Haarlem

Promotiecommissie

Promotores:	Prof. Dr. F. Berto	Universiteit van Amsterdam
	Prof. Dr. A. Betti	Universiteit van Amsterdam
Copromotor:	Dr. P.M. Hawke	Universiteit van Amsterdam
Overige leden:	Dr. L. Incurvati	Universiteit van Amsterdam
	Prof. Dr. G. Priest	City University of New York
	Dr. K. Schulz	Universiteit van Amsterdam
	Prof. Dr. S.J.L. Smets	Universiteit van Amsterdam
	Prof. Dr. B. Vetter	Freie Universität Berlin

Faculteit der Geesteswetenschappen

This dissertation is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement ERC CoG No. 681404).

erc
European Research Council
Established by the European Commission

There is, indeed, a more *mitigated* scepticism or *academical* philosophy, which may be both durable and useful, and which may, in part, be the result of this Pyrrhonism, or *excessive* scepticism, when its undistinguished doubts are, in some measure, corrected by common sense and reflection

Hume, *Enquiry* (XII, III, p. 111)

Table of Contents

Part I: Imagination

Part II: Similarity

Part III: Philosophical Possibilities

Conclusion

Appendices

Acknowledgements

Hume suggested that truth springs from arguments amongst friends. Regardless of whether there are truths in this dissertation, I feel very grateful for the many inspiring discussions I've had, and the many friends I've made, along the way.

First of all, my supervisors Franz Berto, Arianna Betti, and Peter Hawke have helped me a lot in numerous ways. Franz has been my supervisor ever since my Master of Logic thesis and has gotten me into this incredible field of the epistemology of modality, where many of my interests come together. He has been very supportive of my random ideas, interdisciplinary aspirations, and of my academic career in general. I've also worked with Arianna since before the start of this project and she has been a great academic mentor ever since, guiding me in how to connect philosophical and empirical ideas as well as pushing me to write with the greatest conceptual clarity possible. Peter's help has always been invaluable, both in general and pertaining to this dissertation. I feel very lucky to have had a supervisor who is so involved in the topics of this dissertation, who is a great co-author, and most importantly a great friend to have philosophical arguments with (in Rome, Scotland, Amsterdam, and many other places). I hope that our discussions and collaborations will extend beyond the end of this project.

Secondly, I would like to thank Luca Incurvati, Graham Priest, Katrin Schulz, Sonja Smets, and Barbara Vetter for kindly agreeing to serve on my doctorate committee and for reading this dissertation.

Thanks are also due to the other members of the *LoC*-gang. Having such a supportive and friendly academic home has made this whole endeavour seem less insurmountable than it otherwise would have. Our bi-weekly seminars, where I was able to test many ideas in their early stages, were incredibly helpful. Aybüke Özgün is the greatest logician co-author that a philosopher could wish for and Karolina Krzyżanowska is a great sparring partner in all issues concerning conditionals, psychology of reasoning, and how to properly deal with empirical work as a philosopher.

Over the course of this project, I have enjoyed many great discussions that have been inspirational and helped the ideas developed here to higher grounds. I am very grateful to Barbara Vetter and Sonia Roca-Royes, whose work has inspired much of my thinking about modality and the epistemology thereof, for our discussions that have been invaluable and inspiring at different stages of this project. A special thanks to Max Jones for being a great co-author, our discussions about imagination, embodied cognition, and modality have shaped the way I think about the role of imagination in the epistemology of modality. I am much indebted to Chris Badura, Ilaria Canavotto, Martin Lipman, Deb Marber, Becky Millar, Pierre Saint-Germain, Henry Schiller, and Anthi Solaki, for the many discussions I've had with each of them and the feedback they've given on my work over the years. This always resulted in many improvements, the cumulation of which are the chapters presented here. Thanks to Maria Aloni, Elsbeth Brouwer, Jelle Bruinenberg, Catarina Dutilh Novaes, Jorge Ferreira, Dominic Gregory, Levin Hornischer, Savvas Ioannou, Manuel Gustavo Isaac, Drew Johnson, Felipe Morales Carbonell, Ethan Landes, Michela Massimi, Matthew McGrath, Dean McHugh, Vincent Peluce, Thomas Schindler, Margot Strohminger, Ravi Thakral, and Nathan Wildman for the valuable discussions. Thanks to my dad for answering my many questions about statistics, statistical power analyses, and interpreting empirical data. Finally, I am indebted to Brian Rabern, Anders Schoubye, and Jeroen de Ridder for their classes and support, which got me interested in, and ultimately led me to pursue a PhD in, philosophy.

I owe a special thanks to Thom van Gessel. Ever since we started the Master of Logic together, more than six years ago, we have discussed many philosophical, political, and nonsensical ideas, often while enjoying trips and great beers together. We submitted our dissertations within a day of each other and still he found the time to read mine entirely, spotting countless typos, and making many helpful comments.

Not only has this dissertation benefited from many philosophical arguments with friends, the having of friends in general is a *sine qua non* for the completion of this kind of project. I would like to thank all my friends and office-mates both in Amsterdam and during my visit at Arché, St. Andrews. In particular, I would like to thank Anthi, Aybüke, Ilaria, Gianluca, and Thom, their friendship has been very valuable, whether in Amsterdam or on hikes in Edinburgh and Oslo. Thanks are also due to Becky and Henry for all the great times in Amsterdam and Edinburgh; to Stella for all the travels and chess games throughout the UK; and to Jord and Roy for our yearly trips. Thanks to Arjen and Martin for all the great conversations we've had over beers, to Diederik for our morning-coffees, and to Rianne for help with the layout. Finally, a special thanks to Quint and Janine for their incredible support; the numerous coffees, beers, and dinners; and for everything else.

I'm also very much indebted to my parents and sister for their endless support and the countless 'gezellige' evenings with great conversations while playing cards.

Thanks to Lou and Di for providing much needed distraction in the final months of writing. And to Lieke, without whom this would not be possible.

Finally, apologies to Edgar Allan Poe for abusing his title.

Tom Schoonen
Amsterdam, October 2020

Chapter 1

Aims and Assumptions

We have many justified beliefs about what the actual world is like. For example, I justifiably believe that my coffee cup is currently empty, that I am working on my dissertation from home, and I am listening to *Inca Roads* by Frank Zappa. Interestingly, given what is actually the case, we also seem to have many beliefs about what is possibly the case. I could be drinking tea, rather than coffee; I could be working in my office; and I could be listening to different music than I actually am. The justification for beliefs about the actual world can often be explained in terms of our perceptual relations to what is actually the case. It is not obvious that the same can be said about our beliefs about what is possible. An interesting question is, what, if anything, justifies our beliefs about non-actual possibilities?

In this dissertation, I will critically evaluate and propose different *epistemologies of possibility*. In particular, I will discuss *imagination-based* and *similarity-based* theories; two of the main approaches to the epistemology of possibility. The former suggest that imagining something provides *prima facie* justification for the possibility of what we imagined (often under particular conditions of imagination as we will see throughout Part I). The latter, on the other hand, suggest that we can use our beliefs about the actual world and justify beliefs about possibilities that involve relevantly similar objects or situations. In Part II, I discuss potential ways of defining *relevant similarity* and propose a new similarity-based epistemology of possibility.

In this introductory chapter, I will elaborate on some of the basic notions involved in an epistemology of possibility (e.g., what is the kind of possibility that we are interested in? what is justification? etc.). I will do this in the process of explicating three main assumptions that I make for the purposes of this dissertation. These main assumptions will be discussed in Sections 1.1-1.3. I consider an important methodological recommendation that these assumptions give rise to, and which plays an important role throughout this dissertation, in Section 1.2.2. In Section 1.5, I discuss the notion of *justification* and I conclude this introduction by giving an overview of the chapters to come in Section 1.6.

1.1 Objective Modality and Modal Realism

An epistemology of possibility is concerned with our justified beliefs about possibilities. Whenever one presents an epistemology *of something*, there are at least three 'levels' involved in theorising. First of all, there are the objects of the epistemology – i.e., the things justified beliefs about which the epistemology in question aims to explain. Call this the *object level*. Then, there is the *data* that epistemologists themselves should take as their starting point when theorising about the epistemology in question. Call this the *theoretical level*. Finally, there are the ways in which agents, according to the proposed epistemology, are supposed to acquire justified beliefs. Call this the *methodological level*.

In the following three sections, I will address three main assumptions that are made in this dissertation, which correspond to these three questions. In this section, I focus on the first assumption, which corresponds to the object level question for an epistemology of possibility.

1.1.1 Metaphysics of Modality

The epistemology of possibility is a particular case of an *epistemology of modality*. The latter is more general in that it focuses on what, if anything, is the justification for our beliefs in *modal* claims. That is, an epistemology of modality focuses not just on beliefs in possibility claims, for example I believe that I could be listening to different music than I actually am,[1] but also on beliefs in *necessity* claims, e.g., I believe that I am *necessarily* human (i.e., I could not be non-human).[2]

Dummett (1959) remarked that there are two main questions about modality: what is it and how do we 'recognise' it? The former concerns *the metaphysics of modality* and the latter *the epistemology of modality*. The object level question for the epistemology of possibility requires us to at least say something about the metaphysics of modality.

There are many different *kinds* of modality and relatedly many ways in which something can be possible. For example, *epistemic possibility* is possibility given the knowledge or evidence of an agent; *technological possibility* concerns what is possible given our current technologies; logical possibility is possibility following from a

[1] There is an interesting question concerning the distinction between *de re* and *de dicto* possibility claims. For example, 'it is possible that this cup breaks' and 'this cup could possibly break' (where the former is *de dicto* and the latter *de re*). Though there are interesting philosophical issues, especially surrounding the possibility of *de re* possibility claims, I will ignore these and the distinction. I assume that all my examples concern *rigid reference* to the objects involved, so there is no *de re* and *de dicto* distinction between the relevant possibility statements (Fitting & Mendelsohn, 1998, p. 213). I will thus also interchangeably talk of 'it is possible that this object has this property' and 'this object could possibly have this property'.

[2] This example is already theory-laden and potentially controversial, see for example Mackie (2009) for a dissenting voice.

particular logic; *nomic possibility* (or 'nomological' possibility) is what is possible given the laws of nature; et cetera. So, we should narrow down which ones of these we are concerned with before providing an epistemology thereof.

I will focus on possibilities of a particular kind of *factive* modality, namely *alethic* modality. Factive modality are those modalities such that if a proposition is necessary according to that modality, the proposition is true.[3] By focusing on alethic modalities, we *rule out* modalities such as deontic and legal modalities (e.g., the fact that according to some ethical theory you must not kill, does not entail that you do not kill). Within this class of factive modalities, I focus on *alethic modalities*: modalities that do *not* depend on "what any actual or hypothetical agent knows, or believes, or has some other psychological attitude to" (Williamson, 2016b, p. 454).

Assumption 1: *Modal Objectivity*
Focus on justified beliefs in *alethic* possibilities.

That is, we focus on mind-independent, factive modalities. These include, at least, logical modality, nomic modality, causal modality, biological modality, etc., but this rules out *epistemic* modality.[4]

Note on Metaphysical Modality

Philosophers are often interested in a particular kind of alethic modality: *metaphysical modality*. For example, the fact that, given that I am human, I *could not be* anything other than human is a metaphysical necessity. There are, roughly, two ways of characterising what *metaphysical* modality is supposed to be. Either it the modality that captures what follows from and is compatible with one's *essence* (Fine, 1994; Kment, 2014).[5] Or, we define it as the *most inclusive* objective modality (e.g., Van Inwagen, 1998; Hale, 2003; Williamson, 2016b; Strohminger & Yli-Vakkuri, 2018a). An example of this latter characterisation is Williamson (2016b), who defines metaphysical modality as the most inclusive objective modality, such that "metaphysical necessity implies every objective kind of necessity, and dually every objective kind of possibility entails metaphysical possibility" (p. 458). This means that whatever is possible given the laws of nature is also metaphysically possible and whatever is metaphysically necessary is necessary given the laws of nature, but the reverse of these need not hold.[6]

[3] I follow Nolan (2011) in his terminology. Priest (2018, p. 3) calls these kinds of modalities 'veridical'.

[4] See Fine (2002); Nolan (2011); Kment (2014); Williamson (2016b); Kment (2017); and Priest (2018) for some excellent overviews of the different kinds of modalities and the relations between them.

[5] I take this to include both a proper Aristotelian view as well as the Kripkean view (see Kment, 2014 and Priest, 2018 for an overview).

[6] Nolan (2011) provides a nice discussion of such 'absolute'-definition of metaphysical modality and what the acceptance of *impossible* worlds does to such definitions.

Within the epistemology of modality, almost everyone focuses on the epistemology of *metaphysical* modality (Vaidya, 2017).[7] However, some philosophers have expressed sceptical worries about the notion (or the usefulness thereof) of *metaphysical* modality in general (e.g., Putnam, 1990; Nolan, 2011; Priest, 2018; Clarke-Doane, 2019a). For example, Clarke-Doane (2019a,b) has forcefully argued that metaphysical modality is not the most inclusive (or absolute) objective modality.[8] He suggests that we can construct notions of logical modality that satisfy any constraints proposed by the metaphysicians (e.g., Necessity of Identity) and would still be such that they are more inclusive (or absolute) than 'metaphysical' modality. Additionally, Priest (2018) points out that there are no satisfactory arguments for the 'essentialism' route to motivate metaphysical modality being interestingly distinct from (nomo)logical modality.

This dissertation is compatible with both the acceptance of metaphysical modality and with a more sceptical stance towards the usefulness of this notion. I focus mainly on mundane, non-actual possibilities (e.g., that I could be listening to different music than I actually am), which are often nomic possibilities. As Williamson notes, it is easy to see that "if nomic modality is an objective modality, nomic possibility entails metaphysical possibility, the most general type of objective probability" (2016b, p. 486). This means that worries about metaphysical modality do not affect the arguments in this dissertation. If these worries are grounded, it is still of interest to work out an epistemology of nomic possibilities; whereas if these worries turn out to be wrong, the epistemologies presented here also result in justified beliefs about metaphysical possibilities.

Two terminological notes. From now on, I will drop the qualification 'metaphysical' and whenever I talk of 'possibility', 'necessity', 'modality' or the like, I mean *metaphysical* unless otherwise specified. Formally, I will use '$\Diamond\varphi$' to express that it is metaphysically possible that φ (or, equivalently, that it could be that φ). Similarly, I take '$\Box\varphi$' to mean 'it is metaphysically necessary that φ'.

1.2 Modal Mooreanism

Now that we know that we are interested in an epistemology of possibility that focuses on beliefs about alethic possibilities, the question arises where we should start (with theorising) – i.e., the theoretical level question. For example, do we have *any* justified true beliefs about what is possible or not? What should we, as epistemologists, take as our data for theorising about the epistemology of possibility? The *data* available to epistemologists of possibility is highly controversial. In taking

[7] Williamson (2016b, p. 453), Strohminger & Yli-Vakkuri (2018a), and others have argued that we should look at the epistemology of modality more holistically and study the epistemology of metaphysical modality in relation to the epistemology of other objective modalities.

[8] See also Nolan (2011) on whether metaphysical necessity is 'necessity in the widest sense'.

our pre-theoretic judgements about what is (metaphysically) possible to be our data (or at least one of our main sources thereof), there is a worry that "there are a number of possibility claims about which there is much controversy: some philosophers seem to see possibilities where others do not" (Leon, 2017, p. 247). For example, whether or not it is possible that there can be exact physical duplicates of us that lack consciousness is a much debated issue (e.g., Chalmers, 1996; Hill, 1997; Brueckner, 2001; Stoljar, 2007).

In order to address these questions, I follow Leon (2017, p. 247) and assume *Modal Mooreanism*:

Assumption 2: *Modal Mooreanism*
We restrict ourselves to ordinary, Moorean possibility claims as the primary data for the evaluation and construction of theories.

Moorean propositions, in general, are things that are almost university believed by philosophers and non-philosophers, they are the propositions that we justifiably believe, if we have justified beliefs at all, and "they retain their resilience and buoyancy in the face of skeptical worries" (Leon, 2017, p. 247). For example, Moore (1939) famously argued that the premise 'I have hands' is more certain than sceptical arguments against the existence of the external world (see also Lycan, 2019). As Leon (2017) points out, there are also many Moorean *possibility claims* (in addition to the traditional Moorean propositions). For example, we justifiably believe that the glass could break when hit by a ball, I justifiably believe that I could be listening to different music than I actually am, I justifiably believe that the furniture in my room could be arranged differently, et cetera.

I assume that we *do* have justified beliefs in possibilities, namely in these Moorean possibilities, and that this is a phenomenon that an epistemology of possibility should account for. We get around the worry of controversial data points by focusing on Moorean, ordinary possibility claims (e.g., I know that I could be listening to different music than I actually am). Focusing on these basic modal claims allows the epistemologist of possibility to "rely for her theorizing on data of the highest quality, and let the epistemic chips fall where they may" (Leon, 2017, p. 248).

Before I turn to the final assumption, let me discuss two consequences of assuming modal Mooreanism.

1.2.1 Radical (Modal) Scepticism

Radical scepticism is the view that we do not have *any* justified beliefs; a radical sceptic would even contest my claim that I know that I have hands or that I justifiably believe that I am currently in Amsterdam.[9] Radical scepticism is a form of

[9] See Comesaña & Klein (2019) and the relevant entries in Moser (2002); Dancy et al. (2010) for excellent overviews of the debates surrounding scepticism.

global scepticism: we do not have justified beliefs, regardless of the domain of interest. Alternatively, one can adopt *local* scepticism: we do not have justified beliefs *with respect to a particular domain* and we might have justified beliefs outside of this domain. Radical *modal* scepticism is an instance of local scepticism that suggests that we do not have any justified beliefs concerning modal matters.

Accepting Modal Mooreanism allows us to reject radical modal scepticism (and *a fortiori* global scepticism in general). According to modal Mooreanism we have at least *some* justified beliefs concerning possibilities. For example, I justifiably believe that I could be wearing a different shirt than the one that I am actually wearing. I believe that I can leave this room without falling through a black hole. As Hawke puts it, "basic modal claims are somewhat sacrosanct [...] a theory of modal epistemology or modal metaphysics is likely to be viewed with suspicion if it suggests that we are not justified in believing basic modal claims" (2011, p. 360).[10] Note, that rejecting radical allows for the acceptance of more local or moderate forms of modal scepticism. For example, even though I take it that we justifiably believe that Hillary Clinton could have won the 2016 US presidential election, I might still reject the idea that we justifiably believe, or can come to believe, that there could be philosophical zombies (Van Inwagen, 1998; Hawke, 2011). In fact, in this dissertation I conclude that we should adopt a moderate form of modal scepticism: *modal modesty*.[11] We can come to justifiably believe things about what is possible and not, but we should be modest in the range of these kinds of beliefs.

1.2.2 Epistemologies of Possibility and Necessity

Remember from the previous section that the epistemology of possibility is only part of a full-blown epistemology of modality. On top of our beliefs in possibilities, an epistemology of modality also needs to explain our beliefs in necessities – i.e., a complete epistemology of modality also requires an epistemology of necessity. Given the usual interdefinability of possibility and necessity,[12] interesting questions arise about the relation between the epistemology of possibility and the epistemology of necessity. For example, Hale (2003) suggests that there are two *asymmetrical* ap-

[10] Findings from the (developmental) cognitive sciences and the growing literature in the psychology of modality suggest that have beliefs, and reason on the basis of these beliefs, about alethic modality (see Byrne, 2005; Nichols, 2006a; Rafetseder et al., 2010; Gopnik & Walker, 2013; Lane et al., 2016; Phillips & Knobe, 2018; Redshaw et al., 2018; Phillips et al., 2019; Leahy & Carey, 2020). Though such findings would be compatible with radical scepticism, if the sceptics can explain away the evidence that humans seem to reason with and make decisions based on beliefs about what is possible.

[11] The standard label for this view is 'moderate modal scepticism' (Van Inwagen, 1998; Hawke, 2011; Fischer, 2016a; Leon, 2017; Strohminger & Yli-Vakkuri, 2018b); however 'modal modesty' seems to be a much more appropriate label for this view. I will discuss modal modesty in more detail in Chapter 11.

[12] Something is possible just in case its negation is not necessary (in symbols: $\Diamond\varphi \equiv \neg\Box\neg\varphi$) and *vice versa*.

proaches: "*necessity-based* approaches, which treat knowledge of necessities as more fundamental, and *possibility-based* approaches, which accord priority to knowledge of possibilities" (Hale, 2003, pp. 5-6, original emphases). According to such asymmetrical approaches, one focuses on the epistemology of the dominant modality (e.g., possibility) and explains the epistemology of the recessive modality (e.g., necessity) purely in terms of a lack of conflicting dominant claims (e.g., a method of gaining justified beliefs of necessities is also a method of gaining justified beliefs that there are no conflicting possibilities). Alternatively, one might also *reject* either asymmetrical approach and adopt a *symmetrical* approach, where our epistemology of possibility and necessity are (largely) independent of each other (Fischer, 2016a, pp. 76-77). This is sometimes called a *non-uniform approach.*[13]

Focusing on the mundane modal beliefs of ordinary cognitive agents – as per modal Mooreanism – suggests we focus on the epistemology of possibility. For example, consider the following situation:

> A group of young children is playing outside, kicking a ball around in a friendly game of soccer. One of the kids kicks the ball too hard, it bounces off of the curb, and is launched in the direction of a window of one of the neighbours. Luckily, the ball bounced off the window without breaking it, but the children are all very much aware that the ball *could have broken* the window.

The children thus seem to justifiably believe that it is possible for the window to break, but it seems rather far-fetched to suggest that the children do so because they have some beliefs concerning the necessary or essential features of the window or the ball.[14] In particular, statements such as 'property P is an essential property of the ball' do not enjoy the Moorean status that 'The window could break' does. This, I take it, makes the explanation that an asymmetrical *necessity-based* approach gives of our justified beliefs in mundane possibilities (e.g., that the window could break), highly unlikely to be true. However, it is important to stress that the work in this dissertation is compatible with an asymmetrical possibility-based approach as well as a non-uniform approach.[15]

[13] Sonia Roca-Royes is one of the main explicit defenders of a *non-uniform* epistemology of modality (see, e.g., Roca-Royes, 2007, 2017, 2019b, forthcoming). Many others have hinted at something like this, e.g., Strohminger (2015); Fischer (2017a); Leon (2017); Vaidya (2017); see also Mallozzi (2019, sec. 1.4.4) for a brief discussion.

[14] Hawke (2011, 2017) and Roca-Royes (2007, 2017) provide similar arguments to motivate focusing on the epistemology of possibility concerning the modal beliefs of ordinary people.

[15] I believe that the epistemology of modality will ultimately be *non-uniform.* That is, I think that the epistemology of possibility will probably be significantly different from an epistemology of necessity. This means that, in general, I adopt a *symmetric* approach in that I think that not all knowledge of necessities are derived from prior knowledge of possibilities (nor the other way around). However, as I said, every epistemology of possibility discussed in this dissertation is compatible with an asymmetrical possibility-based approach.

Problem of Prior Modal Knowledge

Focusing on an epistemology of possibility comes with a methodological warning. For example, Hale (2003) argued that the 'base class of dominant modal truths' (where these are supposed to be the possibilities such that a possibility-based approach can explain our justified beliefs in them) should be such that "they can be known *without* reliance upon any recessive modality claims" (2003, p. 8, emphasis added). Even within a non-uniform epistemology of modality, when focusing on the epistemology of possibility claims, the epistemology should *not* rely on prior knowledge of necessities.[16]

This methodological recommendation features prominently throughout this dissertation (it will be raised as a serious issue for the theories discussed in Chapters 3, 4, and, to some extent, in Chapter 7):

Problem of Prior Modal Knowledge
Relying on prior knowledge of necessity is a methodological non-starter for epistemologies of possibility.

This methodological warning has been echoed throughout the literature. For example, Roca-Royes (2017) points out that,

> [t]he methodological recommendation that emerges by reflecting on the issue of epistemic priority is as follows: aim at elucidating the [...] possibility knowledge that we have [...] in such a way that success here is *not* parasitic upon success in explaining knowledge of their essential facts. (p. 223, emphasis added)

That is, independently of what you think that correct approach is to the epistemology of modality, when providing an epistemology of possibility one should *not* rely on prior knowledge of necessities (Hill, 2006, p. 230 and Wright, 2018, p. 278).

1.3 Cognitive Plausibility

We now have a better grip on the objects of the justified beliefs that an epistemology of possibility is supposed to explain the justification of – i.e., alethic modality – and the data that we take as our starting point – i.e., ordinary possibility claims. The question that we still need to address for an epistemology of possibility is on the methodological level: in what way do agents come to jusifiably believe things about modality according to the epistemology in question? For example, can we, as epistemologists, suggest that agents have justified beliefs in possibility through

[16] See the discussion in Chapter 3 (Section 3.6) for a more detailed argument on why this is also the case for non-uniform epistemologies of modality.

some special faculty of rational intuition (Bealer, 2002) or modal insight (Fiocco, 2007), or perhaps even by divine intervention (Leftow, 2012).

By providing an epistemology of ordinary mundane claims, we want to explain how it is that we *ordinary humans* can have justified beliefs about those possibilities. So, we should aim to provide a theory that is *cognitively plausible*. That is, whether or not the proposal is such that it describes *how we actually* gain justified modal beliefs, it should at the very least be such that it *could* be the way that we gain modal beliefs. As Roca-Royes (2017, p. 226) points out, the former is ultimately a question for *modal psychology* (e.g., Nichols, 2006a; Phillips & Knobe, 2018; Phillips et al., 2019; Leahy & Carey, 2020), whereas we, as epistemologists of modality, focus on the knowledge- or justification-*conferring* aspect of our epistemological theory (see Section 1.5 below).

Assumption 3: *Cognitive Plausibility*

> A plausible epistemology of possibility "should subsume our capacity to discriminate metaphysical possibilities from metaphysical impossibilities under more general cognitive capacities used in ordinary life."
>
> (Williamson, 2007, p. 136)

The assumption of cognitive plausibility is in line with *non-exceptionalism*, which is refers to the idea that *philosophising* does not require any cognitive capacities beyond those that we, humans, already possess for our ordinary, everyday interaction with the world.[17] As Machery puts it, "the judgments elicited by philosophical cases [...] are warranted, if they are, for the very reason that everyday judgments are warranted, whatever that is" (2017, p. 21). I take it that non-exceptionalism in this sense is a particular instance of the cognitive plausibility assumption, namely, we should rely on "cognitive capacities [that] are not merely a theoretician's dream, but something that we imperfect subjects actually possess" (Balcerak Jackson, 2016, pp. 58-59).[18]

The requirement of cognitive plausibility already allows us to rule out certain approaches to the epistemology of modality. Take for example Chalmers' (2002) *moderate modal rationalism*. According to this theory, conceivability is a guide to possibility, *if* it is considered in a highly idealised way. In particular, conceivability is a guide to possibility if a highly idealised agent – akin to a Laplacian demon – does the conceiving.[19] The worry for a cognitively plausible epistemology of modality would be: "Is conceiving, so understood, a cognitive capacity that we actually have?" (Balcerak Jackson, 2016, p. 57). What good does conceivability do us as

[17] This assumption, together with the 'Objective Modality' assumption, characterise what Vetter (2017, p. 766) calls a 'Williamsonian epistemology of modality'.

[18] See also Vetter, who takes non-exceptionalism to be the claim that "our knowledge of metaphysical modality is *continuous* with our everyday knowledge about the world" (Vetter, 2017, p. 766, original emphasis; see also Williamson, 2016b, p. 487).

[19] See Chapter 2 (Section 2.1.1) for more on conceivability-based epistemologies of modality.

an epistemological tool to gain justification for beliefs about possibilities if it is not within our cognitive capacities to conceive (in the relevant way).[20] Balcerak Jackson (2016) points out that, in general, the assumption of cognitive plausibility rules out epistemologies that rely on highly idealised cognitive capacities such as Chalmers' conceivability.

Accepting cognitive plausibility has further consequences of interest. Let me discuss some of these.

1.3.1 Methodological Naturalism

Given that we focus on a cognitively plausible epistemology, we are concerned with the question of what enables the cognitive processes of ordinary agents to obtain their justificatory role in our beliefs about what is possible. Part of addressing this question involves a careful study of these cognitive processes and their properties. Similarly, attempting to make our epistemology of possibility cognitively plausible involves evaluating it against the backdrop of our best scientific theories of human cognition. That is, I take it that the answer to the methodological level question – i.e., in what way do agents come to justifiably believe things about possibility – should be supported by our best scientific theories.

Phrased differently, I take it that the ways in which agents acquire justified beliefs according to our epistemologies of possibility "should at least be informed and beholden to the results of scientific disciplines" (Goldman, 1994, p. 305). That is, we should adopt *methodological naturalism* (Goldman, 1994; Jenkins, 2013; Nolan, 2017; Rysiew, 2020).[21] Let me stress that this is distinct from, though compatible with, the claim that we have justified beliefs in what is possible by relying on scientific theories.[22] I mean that the proposal of our epistemology of possibility should be in line with what science tells about the cognitive capacities of human beings.

For example, the approaches of, e.g., Williamson (2007) and Machery (2017) are examples of philosophical methods in line with the methodological naturalism intended here. They both provide explanations of our justified beliefs in what is possible based on cognitive capacities of humans that lend themselves to scientific (and sometimes evolutionary) explanations.[23]

[20] This is similar to *The Conditions Question* discussed by Vaidya (2017). "[A method] is useless as a reliable guide to possibility, if it turns out that we are never in the appropriate conditions for [that method] to be reliable" (Vaidya, 2017, p. 99).

[21] I take *methodological* naturalism to be different from *metaphysical* naturalism (Nolan, 2017, p. 8) and focus on the former. From now on, I will often drop the qualification 'methodological', so 'naturalism' should be read as referring to methodological naturalism.

[22] Fischer (2016b, 2017b) provides a very interesting epistemology of modality based on compatibility with our scientific theories. As I mention in Chapter 8 (Section 8.7), I hope to compare and potentially relate Fischer's epistemology of modality with my proposal of that chapter in future work.

[23] Importantly, note that Williamson (2014) explicitly rejects the 'naturalism' label. Naturalists,

1.3.2 Against Modal Rationalism

Methodological naturalists tend to be hostile to epistemologies that posit faculties or methods that are unanswerable to (or even take priority over) the best scientific theories and methods. They typically eschew accounts of knowledge or justification that ignore the limits on human cognition posited by our best scientific theories. For example, "the postulation by philosophers of a special cognitive capacity exclusive to philosophical or quasi-philosophical thinking looks like a scam" (Williamson, 2007, p. 136). Relatedly, and important for our purposes, naturalists are generally suspicious of the rationalistic claim that philosophical knowledge is an *a priori* product of an infallible type of insight, intuition, or reflection that philosophers are specially attuned to. The existence and reliability of such faculties, it might seem, escapes empirical support. As Machery points out,

> It is all too easy to postulate faculties when it suits one's epistemology, one's metaphysics, or one's theology. That there is a faculty of intuition is an empirical claim, which can be only taken seriously if it finds support in our best sciences of the mind—psychology and neuroscience.
>
> (2017, p. 37)

However, he continues, our best sciences "have no place for a faculty of intuition" (ibid.). This is, in a sense, related to the cognitive plausibility worry raised against Chalmers' moderate modal rationalism discussed above. Modal rationalists appeal to methods for acquiring justified beliefs concerning possibilities such as intuition, rational seemings, etc., all of which lack the support of our best empirical sciences of the mind.

Given our assumption of cognitive plausibility, the lack of empirical support for such rationalist methods is a decisive strike against modal rationalism (at least for the purposes of this dissertation).[24] So, aiming for a cognitively plausible epistemology pushes one to modal empiricist, rather than modal rationalist epistemologies.

1.3.3 Fallibilism

Finally, given that I focus on epistemologies based on ordinary human cognitive capacities, the resulting epistemology of possibility will be *fallible.* There are many different formulations of what fallibilism is supposed to be, but the rough idea is that having a justified belief in something is compatible with the *falsity* of the proposition

he complains, tend to equivocate between an unattractively severe position and a harmless but vacuous one. I'm ultimately more interested in the methodological commitments that follow from cognitive plausibility, rather than the choice of label.

[24] Though, as I mentioned, there are modal rationalists who are more receptive to these worries and try to amend their theories accordingly (e.g., Tahko, 2017; Vaidya, 2017; Mallozzi, 2018a).

in question (Leite, 2010, p. 370).[25] Or, phrased in terms of possible defeat, "[a] fallibilist is someone who believes that we can have [. . .] defeasible justification, justification that does not *guarantee* that our beliefs are correct" (Pryor, 2000, p. 518, original emphasis) (see also Cohen, 1988b and Brown, 2018, pp. 1-2). This is also the view that I adopt: if they succeed, the epistemologies of possibility that I discuss provide us with justification for beliefs in possibility claims. Yet, despite this justification, it might turn out that we are wrong – i.e., that what we believed to be possible turns out to be impossible.

I take it that the fallibilism in this dissertation is *motivated* by the methodological assumption of cognitive plausibility. Given that we focus on an epistemology of possibility that relies on our ordinary cognitive capacities and it is uncontroversial that our "generic human cognitive capacity, is fallible" (Williamson, 2016a, p. 177). Holding on to the reliability of these cognitive capacities when concerned with non-actual possibilities, instead of adopting widespread scepticism about our everyday cognitive capacities, suggests that the resulting epistemology (of possibility) is fallible (Williamson, 2007, p. 155 and Balcerak Jackson, 2016, p. 51).

Even though accepting fallibilism suggests that one's epistemology of possibility might sometimes give the wrong predictions (e.g., an agent might be predicted to justifiably believe something to be possible even though it is in fact impossible), one should not fend off all counterexamples to one's theory by claiming fallibilism.

1.4 A Cognitively Plausible Epistemology of Possibility

Let me summarise what we have discussed so far. In this dissertation we focus on the epistemology of (metaphysical) possibility and we take as our starting point (or data) modal Moorean propositions (i.e., ordinary, mundane possibility claims). Vaidya (2016) points out that there are a number of questions that an epistemology of possibility might address. In this dissertation I will evaluate and propose theories that aim to address the *central* question of the field: what is our main (foundational) method for acquiring new justified beliefs about non-actual possibilities?

The epistemology of modality in general is a rapidly growing field, making it impossible to properly discuss all of the available theories. In order to canvas the field a bit, and position this dissertation in it, it will be helpful to consider three jointly inconsistent statements. Giving up any one of these corresponds to a different group of theories within the epistemology of modality. Consider the following three statements (adapted from Roca-Royes, 2007, p. 118).[26]

[25] See Leite (2010) and Dougherty (2011) for excellent overviews of fallibilism, its different formulations, and the discussions surrounding it.

[26] Two things to note here. First of all, this is analogous to the famous *Benacerraf problem* in mathematics. Benacerraf (1973) suggests worries that mathematical facts cannot be known if one

(a) We have justified beliefs about mind-independent possibilities (or modality in general).

(b) Most justified beliefs in mind-independent things are grounded in (perceptual) experiences.

(c) Experientially-based justified beliefs cannot go beyond justified beliefs of mere truths.

The first claim, (a), merely suggests that some things are possible (e.g., it could be sunny instead of raining in Amsterdam now); that these possibilities are *mind-independent* – i.e., they do not depend on thinking minds; and that we have (true) justified beliefs about some of these possibilities. In particular, it suggests that this combines into an interesting fact that needs explaining. The second claim, (b), captures the idea that there needs to be a relation between the cause of one's beliefs about these mind-independent truths and the truth-makers thereof (Benacerraf, 1973, p. 672). For example, we have most of our justified beliefs about the actual world because our senses give us access to what is actually the case. The final claims, (c), is supposed to capture the idea that we "bear no causal relations to the truthmakers for modal claims" (Fischer, 2017a, p. 270), e.g., our experiences cannot teach us anything about truths about non-actual possibilities (e.g., Hale, 2003, p. 1; Roca-Royes, 2007, p. 118).

As Roca-Royes (2007) notes, these three claims are jointly inconsistent, but any two of them taken together are consistent.[27] Considering ways out of this joint inconsistency by giving up either one of these claims allows us to nicely sketch the lay of the land in the epistemology of modality. One might reject (a) and hold that we do in fact not have any justified beliefs in possibilities (note that the claim that we do not have *many* justified beliefs is not enough). The resulting view is that

is a realist about mathematical entities because they seem to be unable to cause the beliefs we have about mathematics (see also Roca-Royes, 2007, p. 118, fn. 2 and Fischer, 2017a, sec. 4).

Secondly, we could pull apart (a) into two separate statements: (i) We have justified beliefs in possibilities and (ii) Modality is mind-independent. If we would do so, we could specify that giving up (i) results in scepticism, whereas giving up (ii) results in a form of 'anti-realism' with regards to modality, such as projectivistm (e.g., Blackburn, 1993), conventionalism (e.g., Sidelle, 1989; Sider, 2013), normativism (e.g., Thomasson, 2013), and expressivism (e.g., Holden, 2014).

[27] Fischer nicely captures the gist of this joint inconsistency.

> Gettier cases show that knowledge is incompatible with (a certain sort of) luck. The most attractive solution to the luck problem requires some causal commerce between the knower and the known, where this interaction explains the knower's epistemic success [(b)]. But if realism about modality is correct, then we bear no causal relations to the truthmakers for modal claims [(c)]. Therefore, if the realist can't provide an alternate solution to the luck problem, she makes our epistemic success [(a)] unintelligible; on her view, it is unclear how we can have any modal knowledge what[so]ever. (2017a, p. 270)

of radical modal scepticism. The main debate in the epistemology of modality is between those who give up (b) and those that give up (c). Giving up (b) results in a form of *modal rationalism*. Conversely, giving up (c) results in forms of *modal empiricism*.[28]

In the beginning of this chapter, I mentioned three levels involved in providing an epistemology of something. As that something in our case is 'possibility', we have the following questions corresponding to the object, theoretical, and methodological level. What is the metaphysical status of the possibilities? What data do we, as epistemologists, appeal to in order to start theorising? And by what means do agents acquire justified beliefs in possibilities? I made three explicit assumptions, one related to each level of the epistemology of possibility. Interestingly, these three assumptions also force our hand in deciding which claim to give up in the jointly inconsistent trio that characterises the epistemology of possibility.

We assumed, as discussed in Section 1.2, that we have some justified beliefs about non-actual possibilities, namely in modal Moorean propositions. Because of this assumption, we reject radical modal scepticism, which means that we *cannot* reject (a) in the inconsistent collection of statements above. Modal rationalists reject (b) and suggest that there are methods of acquiring justified beliefs that do not rely on a causal relation to the truth-maker of those beliefs. Examples of such methods are rational intuitions, seemings, or insights. However, our assumption of cognitive plausibility led us to the rejection of rationalism (Section 1.3.2). So it seems that we cannot (straightforwardly) reject (b).

So, staying true to our assumptions, the only way of getting out of the joint inconsistency would be to reject (c).

1.4.1 Modal Empiricism

As mentioned above, rejecting (c) is something that modal empiricist do, precisely because modal empiricists aim at "finding room for experience to play a larger justificatory role—or even the only role" (Fischer & Leon, 2017a, p. 3). Here, I focus on what Fischer (2017a) calls *liberalised modal empiricism*, which has it that what justifies a modal beliefs is non-modal experiential beliefs *in combination*

[28] See Vaidya (2017) for an excellent overview of rationalism and empiricism in the epistemology of modality and further references to proponents of either side. Note that there are also epistemologies of modality that resist being classified in the rationalist/empiricist dichotomy such as Williamson (2005, 2007). Similarly, there recently has been a push to more 'hybrid' views that appeal to both empiricist and rationalist methods (e.g., Tahko, 2017; Vaidya, 2017; Mallozzi, 2018a). Additionally, those who accept a form of non-uniformism with respect to the epistemology of modality can appeal to empiricist methods for some aspects of their epistemology and to rationalist methods for other parts (e.g., Roca-Royes, 2017 and Roca-Royes, 2019b). Still, even for non-uniform epistemologies of modality, *within* the epistemology of possibility, the discussion of rationalism versus empiricism carries over.

with some ampliative reasoning principles (see also Sjölin Wirling, 2019a, p. 2). Henceforth, I will use ‘modal empiricism’ to refer to liberalised modal empiricism.

Often, modal empiricists appeal to ordinary cognitive capacities that provide justification that, arguably, goes beyond mere truth, such as through imagination or imagery (Kung, 2010; Gregory, 2019), perception (Strohminger, 2015), or similarity reasoning (Hawke, 2011; Roca-Royes, 2017). As mentioned at the beginning of this chapter, within this dissertation I focus on two approaches within the modal empiricist epistemology of possibility: imagination-based (Part I) and similarity-based (Part II) theories. The former is one of the most prominent approaches, though I will argue that it faces some serious difficulties. The latter is what I take to be the most promising approach, though here I will also discuss some initial difficulties such theories have to overcome. This means that I won’t discuss a whole range of other approaches. For example, abduction-based approaches (Biggs, 2011); modalism approaches (see Bueno & Shalkowski, 2014;) perception-based approaches (e.g., Strohminger, 2015); theory-based approaches (Fischer, 2016b, 2017b); and many others.

Let me briefly mention something about one prominent epistemology of modality that is in line with the assumptions of this dissertation that I won’t properly discuss: counterfactual-based approaches (e.g., Williamson, 2005, 2007; Kroedel, 2012, 2017). Williamson, for example, argues for a particular equivalence between counterfactuals and the metaphysical modals of possibility and necessity. Based on this, he argues that we can subsume the epistemology of modality under our epistemology of counterfactuals. The most interesting feature of his account, I take it, is the epistemology of counterfactuals, which relies, amongst other things, on reality-oriented imagination (see Williamson, 2007, ch. 5 and Williamson, 2016a). Throughout the imagination-part of this dissertation, I do often refer to and critically evaluated this aspect of Williamson’s work (in particular in Chapter 4).[29]

A Problem for Modal Empiricism

There are some interesting issues that modal empiricism gives rise to. For example, Fischer (2017a) argues that modal empiricism *leads to* modal scepticism, whereas Vaidya worries whether “the move away from rationalism ultimately require[s] the adoption of a form of anti-realism about modality?” (2017, p. 104). The work in this dissertation goes a long way to alleviating these worries. In particular the proposed empiricist epistemology of modality in Chapter 8 does not fall victim to Vaidya’s worries and, as we will see throughout the dissertation, I take it that we *should* adopt a form of *moderate* modal scepticism (see Chapter 11), which reduces

[29] See Jenkins (2008); Roca-Royes (2011b); Tahko (2012); and Gregory (2017) for more elaborate discussions of Williamson’s counterfactual-based epistemology of modality. Yli-Vakkuri (2013) argues that the imagination-part of Williamson’s epistemology is *not* an essential part for a counterfactual-based epistemology of modality.

some of the pressure of Fischer's worry.

There is another worry, however, that I won't be able to address in this dissertation and it concerns the *integration challenge*. For any given field of inquiry, the intricate relation between its epistemology and metaphysics gives rise to, what Roca-Royes (forthcoming) calls, the *integration requirement*. The integration requirement asks us, for any field of inquiry Φ, to provide "a credible epistemology of Φ-truths that makes justice to the kind of facts Φ-truths are taken to be about" (Roca-Royes, forthcoming, p. 2). For modality in particular, this requirement is not so easy, resulting in an integration *challenge* (Sjölin Wirling, 2019a; Roca-Royes, forthcoming).

The integration challenge is a challenge for both rationalists and empiricists. Modal empiricists, in particular, often raise this issue as problematic for modal rationalists, arguing that rationalist methods fail to properly 'connect' to non-actual possibilities (e.g., Williamson, 2007; Roca-Royes, 2010; Biggs, 2011). However, Sjölin Wirling (2019a,b) argues that modal empiricism is actually "worse off" than the modal rationalist when it comes to the integration challenge (Sjölin Wirling, 2019a, p. 16). She argues that modal empiricists often ignore the modal metaphysics and that, even though we needn't adopt a completely *metaphysics-first* approach (Mallozzi, 2018a), in order for modal empiricism to be evaluated with regards to the integration challenges some modal metaphysics has to be done (idem, p. 7). So, to overcome this worry, Sjölin Wirling argues, "[t]he best option for modal empiricists seems to be to take the IC for modality seriously, and get cracking on the positive story available given her own view, assuming some more particular form of modal realism" (idem, p. 16). Clearly this is a tall order and one that I cannot fulfil here. I agree with Sjölin Wirling that this is a serious issue that ultimately should be worked out, yet I cannot but set it aside for now.[30]

1.4.2 Actuality Principle

Given that I focus on the epistemology of possibility, let me quickly say something about an 'easy' method for gaining justified beliefs concerning possibilities: whatever is actually the case is possible. Call this the *Actuality Principle*: "wherever experience teaches us that p, we may safely reason, *ab esse ad posse*, that it is possible that p" (Hale, 2003, p. 1) (see also Hawke, 2011, p. 360; Nolan, 2011, p. 314; Strohminger, 2015, pp. 372-373, fn. 3; Hanrahan, 2017, sec. 3; Roca-Royes, 2017, p. 229; and many others). This is easy knowledge of possibilities, because it does not require any further epistemological account over and above an epistemology of

[30] As Sjölin Wirling herself notes, there are still some options for a modal empiricist. One attractive option would be to consider a more 'bottom-up' approach to the metaphysics of modality à la Vetter (2015). I completely agree and I think that, for example, the embodied imagination approach developed in Chapter 5 would fit very well with such a potentiality-based modal metaphysics. I hope to develop this link in future work. See also Vetter (2017).

first-order claims about the actual world. Given that I am actually in Amsterdam, it is possible that I am in Amsterdam. So, I know that it is possible that I am in Amsterdam by whatever method I know that I am actually in Amsterdam (plus this simple inference). The fact that it is such easy modal knowledge, also makes it relatively *uninteresting*. This is because the actuality principle doesn't tell us anything beyond what we already knew from the truth about actuality (Hanrahan, 2017, p. 211).[31] As Williamson points out, "the hard question is how far the possible extends *beyond* the actual" (2016b, p. 464, original emphasis).

The aim of this dissertation is to focus on this hard question and evaluate and develop epistemologies of *mere* possibilities – i.e., possibilities that are *non-actual*. Unless otherwise specified, I will use 'possibility' to talk of these mere possibilities. However, as we will see at different points throughout the dissertation, appealing to actuality is often very useful in order to (i) to authenticate certain (prior) possibilities one relies on or (ii) to extrapolate to unactualised possibilities.

1.5 Justification

Throughout this introduction I have talked about 'justified beliefs', as this is a crucial epistemological notion for this dissertation, I will make some preliminary remarks about the way this notion is construed here.

The notion of justification is one of the most crucial of epistemology and we can distinguish between two important questions with regards to it:

- What is justification?
- When are beliefs justified?

These two questions correspond, respectively, with the distinction between meta-epistemological and substantive epistemological questions (e.g., Fumerton, 2002 and Lammenranta, 2004, pp. 467-468). Concerning the epistemology of modality, this dissertation aims to address an instance of the latter: when are we justified to believe modal claims? In order to do so without going into too much detail on the first question, let me make some precursory remarks about justification in general.[32]

First, a clarification concerning the distinction between 'being justified in believing' and 'justifying your belief'. The latter is something that an agent *does* in order to show that their belief is justified, whereas the former "is a state or condition one is in" (Alston, 1985, p. 58). Following most discussions on justification, I focus on

[31] See Hanrahan (2017, sec. 3) for attempts to widen the scope of (something like) the actuality principle.

[32] For excellent overviews of the debates concerning justification see Alston (1985); Fumerton (2002); Lammenranta (2004); Steup & Sosa (2005, Part III); and Steup & Neta (2020, §3). Pappas (1979) is a collection of classical essays on this topic.

questions concerning ‘being justified’ in believing possibility claims. In particular, we focus on a purely *epistemic* interpretation of justification.[33]

I will rely on a very ‘minimalist’ description of what justification is. For the purposes of this dissertation, I take a justified belief to be one that is based on adequate grounds – i.e., something that is believed for the right reasons. For example, if my friends come to believe that they are expecting a child in December and this belief is based on their doctor’s expert testimony, then it is justified; whereas it would not be justified if it is based on fortune-telling through the shape of used coffee grounds or tea leaves. This minimal conception of justification in line with, for example, Goldberg’s minimal characterisation – “a point of widespread agreement: the notion of epistemic justification is a normative notion that applies in virtue of the satisfaction of standards of success in connection to our pursuit of truth (and avoidance of error)” (2015, p. 206) – as well as Fumerton’s – “we might suggest that whatever else epistemic justification for believing some proposition is, it must make *probable* the truth of the proposition believed” (2002, p. 205, original emphasis). Moreover, this minimal characterisation of justification captures the “basic features of the concept that would seem to be common ground” discussed by Alston (1985, pp. 58-59).

1.5.1 The Role of Justification

The above remarks all concerned the meta-epistemological issues concerning what justification is. Let me conclude this discussion on justification by reviewing some issues that bear on the substantial epistemological question of what it is that makes modal beliefs justified

First of all, even though it is important to get clear on the notion of justification, let me point out that I agree with Fischer (2017b) when he says that few (if any) working in the epistemology of modality worry about the distinction between justification and knowledge. The reason for this seems to be that “when it comes to the core questions in the epistemology of modality, little turns on the difference between justification and knowledge” (Fischer, 2017b, p. 6; see also Sjölin Wirling, 2019a, p. 1, fn. 1). I follow suit and throughout this dissertation I will (sloppily) use ‘justifiably believes’ and ‘knows’ interchangeably. Strictly speaking, I focus on *justification* of our modal beliefs, though little turns on this.[34]

Secondly, in this dissertation I will focus on a *process-based epistemology of possibility* (Stuart, 2019). This means that we are concerned with what it is that grounds the justificatory role of the methods that the epistemology of possibility suggests jus-

[33] See Fumerton (2002, p. 205) for a discussion on the distinction between epistemic and nonepistemic justification.

[34] This interchangeable use of ‘justifiably believes’ and ‘knows’ is *not* to say that I think that justification is just that what turns beliefs into knowledge or that justification is knowledge (e.g., Sutton, 2007).

tify beliefs in possibility claims. In our case: what makes that the cognitive capacity under investigation – i.e., imagination and similarity reasoning – are epistemically useful?

Thirdly, I want to get ahead of a possible confusion (and resulting objection). Throughout this dissertation I will use 'ability to provide justification' and 'being epistemically useful' interchangeably. This is because I *focus* on how our beliefs about what is possible are justified. Let me explain. There are different ways in which, e.g., imagination might be epistemically useful: we might use it to *illustrate* a certain point; *explore* the boundaries of our current theories; or *justify* the acceptance (of the possibility) of that which has been imagined.[35] I focus on the *justificatory* role of the cognitive abilities appealed to by the epistemology of possibility (e.g., imagination, similarity reasoning, etc.) and therefore equate 'is epistemically useful' with 'provides justification'. This does not mean that I think that, e.g., imagination might not also be used to explore the boundaries of our current theories (or preconceptions) and that there is some epistemic value and use in this (see Stuart, 2020). It is just that I set this kind of epistemic usefulness aside for the purposes of this dissertation.

Finally, one of the biggest debates concerning the substantive question of justification is the one between internalism versus externalism. Interestingly, a gap in the literature of the epistemology of modality is a careful evaluation of different kinds of theories for their internalist or externalist commitments. The debate between internalism and externalism (with regards to it) is too substantial to review here, so let me just give a brief characterisation of both positions.[36] There are many different characterisations of what internalism is supposed to be, whereas externalism is often defined as the denial of internalism. For example, Alston (1985) discusses three forms of internalism with respect to justification. Justification is such that it is (i) based on mental states of the agent; (ii) internally accessible to the agent; or (iii) (solely) based on other beliefs of the agent. I take (ii) to be the crucial aspect of internalism and that one of its weakest formulations is *weak accessibility (justification) internalism*:

> One has a justified belief that p only if one can become aware by reflection of some essential justifier one then has for p. (Pappas, 2017, §3)

Though I have a personal preference for an externalist perspective (something like two-stage process reliabilism, e.g., Goldman, 1992, ch. 9), most of this dissertation is susceptible to both an internalist or externalist interpretation. For example, in

[35] This is based Machery's (2017, sec. 1.1.2) discussion of the different kinds of uses thought experiments can be put to. See Chapter 9 for an introduction to and a further discussion of the epistemology of thought experiments.

[36] To get some sense of the debate, see Goldman (1979); BonJour (1985); Goldman (1999); Vogel (2000); Kornblith (2001); BonJour & Sosa (2003); Feldman (2014); Greco (2014); and Pappas (2017).

cases where the cognitive process in question might not be internally accessible, internalists could appeal to Wright's (2004; 2014) *entitlement* theory. Discussions between internalists and externalists will not be relevant for the work discussed in this dissertation and I will mostly ignore them. Exceptions are Chapter 5, where I explicitly discuss potential problems for accessibility internalism with regards to a particular theory of imagination, and Chapter 8, where I explicitly discuss both internalist and externalist options for an epistemology of categorisation.

1.6 Chapters: Overview and Origins

This dissertation is divided into three parts. The first two focus on a cognitively plausible epistemology of possibility: the first part in terms of imagination-based epistemologies of possibility (Chapters 3-5) and second part by focusing on similarity-based theories (Chapters 7-8). In the final part, I will turn to the use of possibility statements in philosophy itself and evaluate whether a cognitively plausible epistemology of ordinary possibility statements can provide us with justification for possibility claims that feature in philosophical thought experiments.

In the first part, I concentrate on imagination-based theories, one of the most prominent empiricist approaches to the epistemology of possibility. Roughly, these theories suggest that if one can imagine something (under certain conditions), then one is *prima facie* justified in believing what they imagined to be possible. I first introduce this part by providing an overview of some of the issues in the philosophy and epistemology of imagination. Two main arise: (i) the term 'imagination' is very heterogeneous; there are many, seemingly distinct, cognitive phenomena that we refer to with it (Kind, 2013; Balcerak Jackson, 2018) and (ii) almost everyone in the literature agrees that imagination has to be *restricted* if it is to have any significant epistemological value (Kind, 2016a; Kind & Kung, 2016a; Williamson, 2016a; Balcerak Jackson, 2018). The chapters in this part all concern *different* ways of characterising imagination, all of which have been suggested to play a role in justifying our beliefs in what is possible.

Chapter 3 argues that restricting the linguistic content that features in imagination – especially in theories of representational imagination – does *not* result in a feasible epistemology of possibility. In particular, I will argue that these theories fall victim to the problem of modally bad company: for any pre-theoretically possible situation that one can imagine, there is an impossible situation that relies on similar, restricted linguistic content. I suggest that this problem shows that *if* these theories of imagination can justify our modal beliefs, it is not because of imagination, but because of prior modal knowledge (remember our discussion of Section 1.2.2).

In Chapter 4 I turn to imagination as simulated belief revision (e.g., Nichols & Stich, 2003; Williamson, 2007) and argue that it cannot provide a foundational basis for an epistemology of possibility. I do so by first providing a formal model of pretense imagination and then evaluate the claim that this kind of imagination can justify new beliefs in conditionals that feature in our epistemology of possibility. I conclude that pretense imagination might be used to *expand* our modal beliefs, but it cannot provide justification for beliefs about what is possible without, again, *prior* modal knowledge.

Chapter 5 focuses on theories of imagination as recreating perceptual experiences (Balcerak Jackson, 2018; Gregory, 2019). These theories do not seem to rely on problematic prior modal knowledge. However, I argue that these accounts fall victim to two other problems. In light of these, I will propose a new account of imagination: embodied imagination as sensori-motor simulation. I argue it overcomes these problems. Though this embodied imagination can successfully help us gain justified beliefs about what is possible, the resulting view is limited in the range of possibilities we can justifiably believe because of it.

The conclusions of the first part are largely negative: many interpretations of imagination fail to provide a suitable basis for an epistemology of possibility (with the theories discussed in Chapter 5 as an exception). This motivates looking at a completely different approach in the second part of this dissertation: similarity-based epistemologies of possibility. Very roughly, the idea is that if one knows that an object a actually (and therefore possibly) has property P and object b is relevantly similar to object a, then one can justifiably conclude that it is possible for b to have property P. In the introduction to this part, I elaborate on two *loci classici* of this approach (Hawke, 2011; Roca-Royes, 2017) as well as the general structure of similarity-based reasoning and I stress the importance of the notion of *relevance*.

In Chapter 7, I discuss the literature on similarity reasoning (e.g., Gentner, 1983; Gentner & Markman, 1997) and precisely spell out what is involved in different ways of interpreting 'relevant similarity'. I suggest that one of the most promising interpretations of relevant similarity is as a predictive analogy. I argue that adopting this perspective requires prior knowledge of explicit causal relations, which, depending on one's theory of causality, has severe consequences for a similarity-based epistemology of possibility.

Chapter 8 proposes a similarity-based approach to the epistemology of possibility based on the notion of *kind*. I will develop a technical notion, 'fundamental kind', that will allow us to project possibility claims purely on the basis of observations of the actual world. One key aspect of this theory is our ability to judge two objects to be of the same kind and I will present a variety of potential epistemological explanations based on empirical data from cognitive and developmental psychology. I argue that this results in a similarity-based epistemology of possibility that is knowledge-conferring and does not rely on problematic prior knowledge.

In the third part of the dissertation, I turn to the role that possibility statements play in philosophy. In particular, I discuss the use of *thought experiments* in philosophy. Chapter 9 explores the recent debate concerning the analysis of thought experiments (e.g., Williamson, 2007; Geddes, 2017). I strengthen and expand one of the main objections against the Williamsonian analysis of thought experiments: the problem of deviant defeat. I propose a new solution to this problem, which emphasises that one of the most important open questions in the epistemology of thought experiments is: what is it that justifies agents to believe philosophically interesting possibility claims? This stresses the need for an epistemology of possibility that is suitable for justifying philosophically interesting possibilities (i.e., those used as in thought experiments).

In Chapter 10, I discuss philosophers from the experimental philosophy tradition, who have expressed radical scepticism about our ability to judge, based on our ordinary cognitive capacities, whether philosophically interesting hypothetical cases are possible (Machery, 2017). I discuss the arguments that these theorists give and argue that their *radical* scepticism is unfounded: there are *some* philosophically interesting cases that we can justifiably believe to be possible. This significantly undermines Machery's pessimistic inductive argument in favour of completely rejecting the use of thought experiments. However, despite the rebuttal of this radical claim, a more modest reformulation of Machery's argument remains.

I conclude in Chapter 11. I start by summarising the findings of this dissertation and identify an important theme that runs throughout this dissertation: *modal modesty*. I highlight different varieties of modal modesty, different motivations for it, and some possible consequences of accepting modal modesty.

How to Read this Dissertation

Apart from being read in its entirety from beginning to end, the parts can be read independently of each other. For example, readers interested in similarity-based approaches can read Part II without having to first have to read through Part I. The same goes for the other two parts. Readers familiar with imagination-based and similarity-based approaches to the epistemology of possibility, might skip the introductory chapters to the respective parts. Finally, though the chapters within the imagination-part (Part I) can be read independently of each other, it is recommended to read the chapters within Part II and Part III in order.

Origin of the Material

Some chapters in this dissertation are based on previous work in the form of articles. Here I'll explain which chapters rely on these articles. In case the previous work is co-authored, all authors contributed equally.

- Chapter 3 is based on:

  Schoonen, T. (2020). The Problem of Modally Bad Company. *Res Philosophica*, forthcoming.

  The Appendix to this chapter (Appendix A) is based on parts of:

  Berto, F. & Schoonen, T. (2018). Conceivability and possibility: some dilemmas for Humeans. *Synthese, 195*(6), 2697-2715,

- Sections 4.1-4.3 of Chapter 4, as well as Appendix B, are based on:

  Özgün, A. & Schoonen, T. (in preparation). Modelling Pretense-Imagination over Time.

  The remainder of this chapter (Sections 4.4-4.8) is based on:

  Schoonen, T. (under review). A Note on the Epistemological Value of Pretense-Imagination.

- Chapter 5 is based on:

  Jones, M. & Schoonen, T. (in preparation). Putting Knowledge from Imagination on firmer Grounds.

- Chapter 9 is based on:

  Hawke, P. & Schoonen, T. (in preparation). Gettier Reasoning and the Problem of Defeat.

- Chapter 10 is based on:

  Hawke, P. & Schoonen, T. (2020). Are Gettier Cases Disturbing? *Philosophical Studies*, forthcoming.

Part I

Imagination

Chapter 2

Introduction to Imagination-based Theories

A recent article in The New Yorker, titled 'The Coronavirus is Rewriting our Imaginations,' opens with the sentence 'What felt impossible has become thinkable' (Robinson, 2020). When discussing the effects the coronavirus has on the structure of our society and the 'opportunity' this allows us to make a radical shift in that structure, Robinson puts imagination at the centre stage. Experiencing this current pandemic makes our imaginings about such events more precise, allowing us to think about situations that previously seemed impossible.

> Imagine pandemics deadlier than the coronavirus. These events, and others like them, are easier to imagine now than they were back in January, when they were the stuff of dystopian science fiction. (ibid.)

News reports and ordinary conversations are rife with this, almost automatic, interchangeability between talk of imaginability and possibility. Similarly, the link between imagination and possibility has an impressive philosophical record. Even though some philosophers are *pessimistic* about such a link (e.g., Mill, 1882; Putnam, 1973), there "runs a certain schizophrenia" through philosophy "in which, the theoretical worries forgotten, conceivability evidence is accepted without qualm or question" (Yablo, 1993, p. 2).

In this introduction, I will set up the discussion concerning *imagination-based epistemologies of possibility.* I will discuss some philosophical issues surrounding imagination; issues related to the epistemic use of imagination; worries that arise due to Kripke-Putnam *a posteriori* impossibilities; and argue against accepting an error-theory with regards to imagination. This sets the stage for the imagination-based epistemologies of possibility that will be discussed in detail in the following chapters.

2.1 Philosophy of Imagination

As the above quote from Yablo suggests, philosophers *not* working on questions surrounding the epistemology of modality often accept evidence from imagination without question (as also evidenced by the use of thought experiments; see Chapter 9). The idea, very roughly, is that when we are able to imagine something – say that I write this using pen and paper – then that something must be possible. Or, to phrase it slightly more cautiously, imagining a situation provides us with *prima facie* justification that that situation is possible. Even though this picture enjoys some intuitive appeal, there are a number of questions surrounding the nature of imagination that need to be addressed in order to justify this reliance on imagination.

Over the last twenty years, research into imagination itself has flourished and, consequently, influenced the debate on imagination-based epistemologies (of modality).[1] I will discuss (issues concerning) the epistemological value of imagination in more detail in the following sections. Here, I briefly want to discuss the notion of conceivability; the heterogeneity of imagination; and an initial taxonomy of it.

2.1.1 What is Imagination?

The usage of the term 'imagination' itself is already very heterogeneous – i.e., there seem to be many distinct cognitive phenomena that we refer to with it (Strevenson, 2003; Kind, 2013; Balcerak Jackson, 2018). However, before we discuss a taxonomy of imagination, let me first briefly say a few words on distinguishing it from some close cousins: imagination is closely related to *supposition* and *conceiving*. As Balcerak Jackson (2016, sec. 2) points out, these three attitudes share two features that might make it tempting to group them together. First of all, they all concern *thinking about hypothetical situations*. Whether I ask you to imagine, suppose, or conceive something, in most cases I will ask you this when that thing is a "merely hypothetical" object or situation (Balcerak Jackson, 2016, p. 44). Secondly, all of these attitudes are under our voluntary control: we decide when and what we want to imagine, suppose, or conceive. Balcerak Jackson points out that these commonalities encourage "a natural pre-theoretical assumption that they [i.e., imagination, supposition, and conceivability] are instances of the same basic cognitive capacity" (ibid.). However, she forcefully argues that despite this, imagination, supposition, and conceivability are significantly distinct. Supposition is mostly easily distinguished from the others, as it is much weaker than imagination/conceivability in that it requires less commitment of the agent to the proposition in question. As an example, we might not be able to imagine or conceive of a situation where Stalnaker

[1] Some classical works are Walton (1990); White (1990); Harris (2000); Currie & Ravenscroft (2002); Gendler & Hawthorne (2002a); and Nichols & Stich (2003). See Kind (2016c) and Liao & Gendler (2019) for an overview of the literature and further references. See Gendler & Hawthorne (2002b, sec. 2); Kind (2016c, Part 1); and Kind & Kung (2016a, sec. 2) for overviews of imagination throughout the history of philosophy.

is the smallest prime number (e.g., due to conceptual incoherence; see Yablo, 2002); but "we have no trouble supposing that Stalnaker is the smallest prime number, [. . .], our ability to do so is crucial for our ability to engage in *reductio* reasoning" (Balcerak Jackson, 2016, p. 53). So, even though something like 'temporary acceptance' is characteristic of imagination, conceivability, and supposition, only the former two seem to require something in addition to it.[2]

The distinction between conceivability and imagination is of more interest, as much of the literature in the epistemology of modality (in particular the pre-2005 literature) is phrased in terms of the former (e.g., Van Cleve, 1983; Yablo, 1993; Tidman, 1994; Chalmers, 2002).

Consider some definitions people have given of what conceivability might be in terms of 'modal imagination':

> p is conceivable for me if I can imagine a world that I take to verify p. (Yablo, 1993, p. 29)

and similarly,[3]

> We might say that in these cases, one can *modally imagine* that P. One modally imagines that P if one modally imagines a world that verifies P, or a situation that verifies P. (Chalmers, 2002, p. 151)

These definitions might make a deflationary account of conceivability in terms of imagination tempting; raising the question whether "there still is a place for a distinctive cognitive capacity of conceiving" (Balcerak Jackson, 2016, p. 54).

However, as Balcerak Jackson argues, it is likely that, in particular, Chalmers holds that conceiving is a cognitive capacity significantly distinct from imagination.[4] Chalmers' notion of conceiving can be thought of as "simulating belief" for "ideally rational believers with unlimited reasoning capacities" (Balcerak Jackson, 2016, p.

[2] See Arcangeli (2018) for an elaborate discussion of what supposition is and its relations to imagination. According to her, supposition is akin to what I call 'pretense imagination' (see Chapter 4).

[3] I say 'similarly,' but there is a significant difference between Yablo's notion of conceivability and that of Chalmers. Yablo presents, what is called, an *epistemic account* of conceivability. It "is epistemic because it is relativized, on the one hand, to S's state of knowledge and, on the other, to S's conceptual resources plus rational capacities. It is, therefore, *subject-relative*" (Roca-Royes, 2011a, p. 24, original emphasis). Chalmers' notion of conceivability, on the other hand, is *non-epistemic*. "Non-epistemic notions use[, as we will see,] *ideal conceivers*" (ibid., original emphasis). Here I take conceivability to be a notion that concerns *idealised* agents à la Chalmers. I suggest that Yablo's notion is related more to the QALC imagination theories discussed in Chapter 3.

[4] This is because for Balckerak Jackson, imagination crucially involves *phenomenal* perspective taking (Balcerak Jackson, 2016, sec. 3 and Balcerak Jackson, 2018). If you think that 'simulated rational belief revision' counts as imagination (see Chapter 4), then a non-idealised version of Chalmers' conceiving might still be counted as a special instance of imagination.

56).[5] Thus understood, and understanding imagination as essentially involving a *phenomenal* aspect, conceiving is distinct from imagination.[6]

For our purposes, I will set the notion of conceivability that concerns *ideal* agents aside. I will focus exclusively on imagination, but I take it that simulated belief revision is also a form of imagination. Chalmers' notion of conceivability surely deserves close study in its own right (which it has received already, see references in footnote 5), but remember that we are concerned with explaining how we, ordinary human beings, gain knowledge of possibilities. Given that it is not at all obvious that this notion of conceivability is "a cognitive capacity we actually have," we can set it aside for the purposes of this dissertation (Balcerak Jackson, 2016, p. 57; see also Worley, 2003 and Roca-Royes, 2011a).

The Heterogeneity of Imagination

Having distinguished imagination from its close cousins supposition and conceivability, we now turn to imagination itself, where we are still faced with the question: what is imagination? Kind and Kung sketch the state of the art best when they point out that

> Anyone coming to the imagination literature for the first time would undoubtedly be frustrated by the lack of a clear explanation of the mental activity being talked about. The problem is not simply that philosophers give different theoretical treatments of imagination but rather that there doesn't even seem to be consensus about what the phenomenon under discussion is. (2016a, p. 3)

Defining what imagination is turns out to be very difficult. In fact, in her introduction to *The Routledge Handbook of Philosophy of Imagination*, Kind writes that "this question has proved remarkably difficult to answer – so much so, in fact, that many authors, including many of the authors in this collection, explicitly refrain from even trying to do so" (2016b, p. 1). One of the reasons why this might prove so difficult is because of the fact that it is very likely that any list of issues that imagination supposedly plays a role in consists of many varied phenomena and continues to grow (Kind, 2013, p. 141). Philosophers suggest that imagination plays a role in our aesthetic judgements; our engagement with fiction; creativity; pretense; action planning; counterfactual reasoning; empathy; thought experiments;

[5] See Chalmers (1996, 1999, 2002); and Chalmers (2010, ch. 6) for his original work on conceivability as a guide to possibility and Brueckner (2001); Gendler & Hawthorne (2002b); Worley (2003); Stoljar (2007); Roca-Royes (2011a); Vaidya (2016); Balcerak Jackson (2016); Strohminger & Yli-Vakkuri (2017); and Mallozzi (2018b) for discussions thereof.

[6] Though many use 'conceivability' interchangeably with 'imagination,' without specifying what they take this to refer to. As just one example, "I take conceiving and imagining to be the same attitude; 'imagining' and 'conceiving' will be used inter-changeably" (Lam, 2017, p. 2156).

epistemology of modality; mindreading; scientific modelling; et cetera. However, "[i]nsofar as philosophers have invoked imagination to explain these very varied activities, they have not always had the same sort of mental activity in mind" (Kind, 2013, p. 143). In fact, Kind forcefully argues that, even if we just focus on the role of imagination in our engagement with fiction, pretense, mindreading, and the epistemology of modality, there is no single cognitive capacity that *can* fulfil these roles. She argues that the respective requirements on a cognitive capacity to play these roles are in conflict with one another and concludes that "there is nothing about the imagination itself that allows it to play all the different explanatory roles that it has been assigned" (Kind, 2013, p. 154).[7]

Instead of proposing all sorts of confusing technical terms to distinguish all these different senses of 'imagination,' I will just use 'imagination' throughout this dissertation. I will specify in each chapter how the term is understood in that chapter and I want to stress that at no point I am suggesting that any of the interpretations that I use of imagination are the *only* or the *correct* way of looking at imagination. I take imagination to be a versatile cognitive capacity, often involving a combination of the kinds of imagination that I discuss over the course of this dissertation, and agree that it is likely not a single, uniform capacity that is the same capacity that we appeal to in different contexts. I agree with Van Leeuwen when he says that "I think we should take [. . .] 'imagination' to refer to a *capacity* and not a *faculty*, since faculty seems to imply a unified, autonomous, specialized mental system – a 'module,' so to speak" (2013, p. 223, original emphases).[8]

Despite this heterogeneity, there is some consensus on a minimal taxonomy of different acts of imagination that almost all theorists agree on (Kind, 2016b). The distinction is between three 'kinds' of imaginative acts: propositional imagination; sensory imagination; and experiential imagination. I'll briefly discuss each in turn.

Propositional imagination is imagining *that* φ. For example, I can imagine that there is a tiger behind the curtain; I can imagine that Mark Twain is playing basketball; and I can imagine that I am at a tea party. In each case an agent has a particular cognitive attitude (imagination) to a particular proposition.[9] Propositional imagining is supposed to be opposed to *objectual imagination*, where "a subject bears an imagination relation to an object or an event [. . .] rather than to a proposition" (Balcerak Jackson, 2018, p. 210).

[7] Note that Kind's arguments are based on the assumption of a *recreativist* approach to imagination in order to explain imagination's role in mindreading. Though I agree that it is likely that this is the best theory of imagination to play that role, others might disagree (e.g., Carruthers, 1996; Carruthers & Smith, 1996).

[8] Throughout this introduction, use of unqualified 'imagination' refers to imagination in any of its forms. Though, as I said before, in the following chapters the use of unqualified 'imagination' refers to the kinds of imagination under discussion in that particular chapter.

[9] I take it that propositional imagination is at least *attitude imagining* and sometimes also *constructive imagining* in Van Leeuwen's (2013) terminology.

Sensory imagination is imagination that includes some form of qualitative content (e.g., imagistic, auditory, olfactory, etc.), as opposed to purely propositional representations (see Van Leeuwen's 2013 discussion of imagistic imagination). Imagining seeing a particular shade of red has a particular qualitative content, whereas imagining how your (political) supporters would react if you voted for gun control might not be accompanied by any perceptual content, in fact such an imagining "need not involve mental imagery" (Williamson, 2016a, p. 117). Mental imagery is the most often used example of sensory imagination (see Van Leeuwen, 2013; Kind, 2016b; Balcerak Jackson, 2018; Macpherson & Dorsch, 2018; Gregory, 2019).

Finally, experiential imaginings are imaginings 'from the inside'. That is, when we imagine seeing a particular shade of red, there is something what it is like to imagine seeing it. "When we're engaged in experiential imagining, we project ourselves into an imagined situation and image the experiences – visual, auditory, emotional, and so on – that we would have" (Kind, 2016b, p. 5). Sensory imagination is often, though not always, accompanied with experiential imagination. Balcerak Jackson, when talking about appearance-based imagination (which I will discuss in Chapter 5), describes it as follows:

> [T]he basic idea is that it is the nature and function of [experiential] imagination to take up various aspects of the phenomenal character and the content of the corresponding actual or non-actual perceptual experiences of actual or non-actual subjects in order to create relevantly similar experiential states. (2018, p. 218)

Given this taxonomy of different kinds of imagination, one might wonder which, if any, of these kinds of imagination is most suitable to play a role in the epistemology of possibility. Consider the case of propositional imagination. There are plenty of propositional contents that represent impossibilities: I can imagine that Frank Zappa is my father, that my cats are cleverly disguised robots, and that unicorns walk the streets of St. Andrews. Similarly for sensory imagination, consider pictorial representations such as the Penrose Triangle, the Impossible Trident, or Escher's famous 'Ascending and Descending'. All of these pictorially represent impossibilities.[10] So it seems that imagination, on most conceptions, can represent impossibilities, how then could it be a suitable basis for an epistemology of possibility?[11]

[10] Though not everyone agrees (e.g., Sorensen, 2002; see Chapter 3, footnote 13).

[11] I intentionally did not mention experiential imaginings representing impossibilities. Chapter 5 explicitly discusses a form of experiential imagination, its modal epistemological features, and how it relates to what we will discuss next: the Puzzle of Imaginative Use.

2.2 The Puzzle of Imaginative Use

In general, there seem to be two, seemingly opposing, uses to which we put our imaginative capacities. On the one had, we seem to use imagination in order to gain new (justified) beliefs (call this the *instructive use*): we imagine moving the couch through the door before starting the heavy lifting; we imagine how the phrasing of our comments might affect our colleague's feelings; we imagine whether or not the stroller fits into the trunk of a car we might want to buy. On the other hand, we can imagine the most fanciful things (call this the *transcendent use*): we imagine Alice falling down a rabbit hole; we imagine Dr. Jekyll being distinct from Mr. Hyde; and we imagine Frank Zappa being our father and imagine what that would be like. Kind and Kung, who coined the labels for these two uses of imagination, point out that these two uses seem incompatible and, thus, give rise to a puzzle.

> As the examples suggest, imagination is put to two distinct and seemingly incompatible kinds of uses. [...] But how can a single mental activity successfully be put to both uses? How can the same mental activity that allows us to fly completely free of reality also teach us something about it? This puzzle—what we'll call the *puzzle of imaginative use*—has received surprisingly scant attention in philosophical discussions of imagination. (2016a, p. 1, original emphasis)

An initial response to the puzzle of imaginative use might be to appeal to the heterogeneity of imagination discussed above. The idea would be that different kinds of imagination play a role in the instructive use of imagination and in the transcendent use of it.

> The equivocation solution to the puzzle of imaginative use derives support from this apparent lack of consensus about what imagination is. According to this proposed solution, the term 'imagination' is equivocal; there are (at least) two different senses of it. Thus, one kind of imagination—call it imagination$_T$—has transcendent use, while another kind of imagination—call it imagination$_I$—is responsible for the instructive use. (idem, p 4)

However, Kind and Kung argue that the distinctions between imagination (as discussed above) *cannot* account for the different uses of imagination. And although they do not provide an argument that there cannot be different kinds of imaginings that line up with the distinction between the instructive and transcendent use, they "are skeptical that such a distinction could be found" (idem, p. 5). In particular, they argue that such an equivocation solution cannot account for the fact that "the power of imagination to transcend the world seems directly *continuous with* its power to teach us about the world. [...] [S]uch continuity remains entirely inexplicable" on an equivocation solution to the problem of imaginative use (ibid., original emphasis).

The solution that Kind and Kung propose is one that most philosophers of imagination agree on: in order to explain the instructive use of imagination (i.e., for it to be epistemically useful) it has to be *restricted*.

2.2.1 Restricting for Epistemic Value

What are the conditions under which imagination is epistemically useful? That is, when can imagination provide us with justification? These kinds of questions are at the centre of recent debates in the epistemology of imagination. One thing that most agree on is that in order for imagination to be epistemically useful, it has to be restricted (see Kind, 2016a,b; Williamson, 2016a; Balcerak Jackson, 2018).[12] In terms of the puzzle of imaginative use:

> [T]he key to solving the puzzle is to acknowledge and explain how our expansive powers of imagination can be *reined in*. When there are *constraints* on imagination, either architectural constraints or constraints that we can willingly impose, and when these constraints ground imagination in the real world in the right way, imagination can help us discover truths about the real world.
>
> (Kind & Kung, 2016a, p. 2, original emphases)

The idea, roughly, is that *unrestricted* imagination allows us to imagine the most fanciful situations; but when we *restrict* imagination it can (potentially) provide us with justification – i.e., it might be instructive. Let me stress that in the puzzle of imaginative use, as discussed by Kind & Kung (2016a), the instructive use of imagination is aimed at knowledge of *actuality* (see also Kind, 2016a). However, we focus on knowledge of *non-actual possibilities*. Much of the discussion carries over: we seem to be able to imagine many impossibilities, but imagination also seems to provide us with justification for beliefs about what might have been the case.

Balcerak Jackson (2018, sec. 3) discusses a closely related puzzle in relation to imagination-based epistemologies of modality. The puzzle she discusses is, what she calls, the *Up-To-Us Challenge*.

[12] See Stuart (2020) for an opposing view. He argues that it is sometimes the *lack* of restrictions that makes imagination epistemically useful. However, what he considers as 'epistemically useful' is significantly different from what we are considering. On his account, it is the *exploratory* role of imagination that is important, whereas we are concerned with the justification for certain beliefs. To talk with Williamson (2016a, p. 115), Stuart seems to argue that we should not dismiss imagination's role in the context of discovery, as this can also be 'epistemically useful'. Whereas, in this dissertation, I focus on the context of justification (see the discussion on justification from Chapter 1, Section 1.5), in particular, in this part, on imagination's ability to *provide justification* for our beliefs in possibility claims. All agree that in that case, imagination has to be restricted.

> Imaginings are under our voluntary control. If imaginings are under our voluntary control then what we imagine is determined by what we want to imagine rather than by how things are. In a slogan: imaginings are up to us. Therefore, imaginings cannot teach us about anything, or at least not about anything that we didn't already know.
>
> (Balcerak Jackson, 2018, p. 212, footnote removed)

In relation to the epistemology of possibility, combining the puzzle of imaginative use with the Up-To-Us challenge results in the following worry: 'willingly imposed' constraints on imagination cannot teach us anything we didn't already know.[13] Balcerak Jackson forcefully argues that the only kinds of restriction that can overcome this worry are *inherent restrictions* and that *recreativist* accounts of imagination are (inherently) restricted in the right way.[14] On such recreativists accounts, imagination is restricted "in virtue of being by their very nature *derived from* or parasitic on" the cognitive capacity that they recreate (Balcerak Jackson, 2018, p. 221, original emphasis). The idea is, roughly, that if imagination recreates, e.g., perceptual experiences and these are inherently restricted by our neuro-physiological make-up, then the resulting imagination inherits these restrictions.

In Chapter 3, I discuss an account of imagination that is *not* recreativist and that, in a sense, relies on *willingly imposed constraints* in order to provide an imagination-based epistemology of possibility. The conclusion of that chapter is in line with Balcearak Jackson's: this kind of imagination can only provide a successful epistemology of possibility by relying on problematic prior modal knowledge – i.e., it relies on modal knowledge we already have. In Chapters 4 and 5, I will focus on recreativist accounts of imagination (Chapter 5 in particular focuses on the justificatory role of chosen versus unchosen constraints).

2.3 Kripke-Putnam *A Posteriori* Impossibilities

Kripke (1980) famously argued for the distinction between *aprioricity* and *necessity*. He points out that "they are dealing with two different domains, two different areas, the epistemological and the metaphysical" (1980, p. 36). The pre-Kripke, traditional conception was that all contingent truths are *a posteriori* and all necessary

[13] The distinction between *chosen* and *unchosen* constraints (or 'willingly imposed' and 'architectural') is rarely considered in the literature and will be discussed elaborately in Chapter 5.

[14] I follow Balcerak Jackson in using the term 'recreativist' instead of 'simulationist' as the latter has too many connotations (e.g., simulation of theory of mind). As Pezzulo and Castelfranchi point out, "[t]he term 'simulation' is used ambiguously in the literature, in several contexts" (2009, p. 561). I therefore prefer the label 'recreativist', however, I take it that many 'simulationist' approaches, such as Currie & Ravenscroft (2002), Nichols & Stich (2003), and Goldman (2006) also are recreativist accounts (even though they might use their recreativist account of imagination for simulationist theories of mind as well).

truths *a priori*.[15] The idea is intuitive: if something is necessary, it is true in all situations, so I do not need any empirical information of what the world is like to be able to come to know it. Conversely, if something could have been different, then I need to look at what the world is like in order to determine whether it is true. However, Kripke showed that these concepts differ *extensionally*; that there are *a posteriori necessities* and *a priori contingencies*. Famous examples of the former include 'Hesperus is Phosphorus' and 'Water is H_2O' and examples of the latter include 'This stick [pointing to the meter-stick] is one meter long' and 'Julius invented the zip' (where we fix the referent of 'Julius' to be whomever invented the zip).

The existence of *a posteriori* necessities in particular is problematic for (imagination-based) epistemologies of possibilities.[16] To see this, consider a characterisation of *a posteriori* necessities as a two-step deductive process:

(1) It is argued that if some fact is true, it is necessarily so ($\varphi \to \Box\varphi$).

(2) By empirical investigation, the relevant fact turns out to be true (φ).

(C) By deduction (from 1 & 2), we conclude that that fact is necessarily true ($\Box\varphi$). (see Yablo, 1993; Hill, 1997)

Before we perform the relevant empirical investigation pertaining to (**2**), we seem perfectly capable of imagining the negation of an *a posteriori* necessity; in fact, these were often believed to be true (e.g., the ancient Greeks believed Hesperus to be distinct from Phosphorus). The acceptance of *a posteriori* necessities thus gives rise to a major problem for imagination-based epistemologies of possibility, for if we can imagine such *impossibilities*, how can imagination then justify our beliefs in what is possible? In particular, it raises the question whether or not "these cases [can] be cordoned off in a principled way, so that one can explain the failure of [imagination] in particular cases while maintaining the general reliability of the practice described" or whether "nothing systematic [can] be said in this regard" (Gendler & Hawthorne, 2002b, p. 10).

Let us call cases of *a posteriori* necessities/impossibilities: *Kripke-Putnam cases*. Putnam (1973, 1975), the other main contributor to the establishment of *a posteriori* necessities, thought these problems to be insurmountable for imagination-based epistemologies of possibility (phrased in terms of 'conceivability'):[17]

> [W]e can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H_2O. In

[15] The 'traditional' conception includes the positivist and Kantian view on these matters (Gendler & Hawthorne, 2002b, sec. 3 and Vaidya, 2016, §1.1).

[16] For some excellent discussion of the modal epistemological worries pertaining to Kripke's work see: Kripke (1980); Yablo (1993); Hill (1997); Gendler & Hawthorne (2002b, sec. 3); and Vaidya (2016, §1.1).

[17] Putnam (1990) later distanced himself from the strict *metaphysical* interpretation of these *a posteriori* necessities.

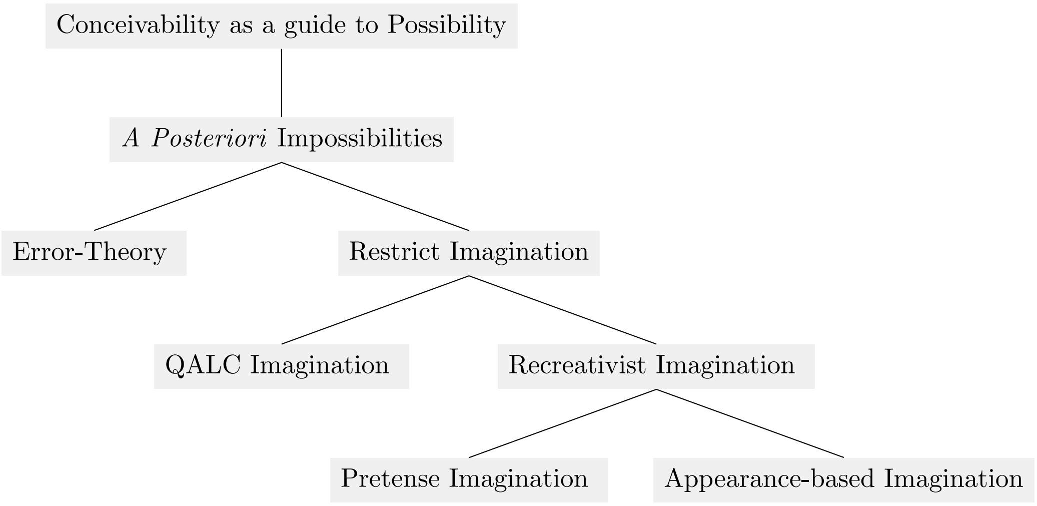

Conceivability as a guide to Possibility
A Posteriori Impossibilities
Error-Theory
Restrict Imagination
QALC Imagination
Recreativist Imagination
Pretense Imagination
Appearance-based Imagination

Figure 2.1: *Responses to Kripke-Putnam cases.*

> that sense, it is conceivable that water isn't H_2O. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.
> (1975, p. 151, original emphasis)

Yet most epistemologists of modality take the issue of the Kripke-Putnam cases as a starting point, rather than giving up in the face of them. Figure 2.1 represents a number of responses to the Kripke-Putnam cases that one finds in the literature.[18] Many propose that the restrictions we impose on imagination ought to be such that they rule out Kripke-Putnam cases. As Byrne (2007) puts it, "imaginability is a guide to possibility only if Kripkean impossibilities are unimaginable" (p. 130). For example, those who focus on the representational aspect of imagination suggest to restrict the linguistic content allowed in imaginings that feature in our epistemology of possibility. I will discuss these theories in Chapter 3 (QALC Imagination). Others, as we saw above, suggest that imagination is inherently restricted by recreating cognitive capacities that themselves are inherently restricted (by, e.g., our neuro-physiological make-up). Some suggest that imagination recreates our ability of rational belief revision, while others suggest that it recreates our perceptual machinery. I will discuss the former in Chapter 4 (Pretense Imagination) and the latter in Chapter 5 (Appearance-based Imagination).

There is another option that suggests we do *not* need to restrict our imagination.

[18] Note that I only focus on those responses relevant for the project of this part of the dissertation. For example, I do not discuss, what Strohminger & Yli-Vakkuri (2017) call, *Two-Factor Views*. "According to these views, *a posteriori* modal knowledge can always be 'factorized' into a modal component that is *a priori* and a non-modal component that is not" (idem, p. 829). Examples of such views include Casullo (2010); Hale (2013); and Mallozzi (2018a). I do not discuss these theories as they are (often) epistemologies of necessity, whereas I focus on epistemologies of possibility (though I will briefly discuss Mallozzi's view in Chapter 8, Section 8.7).

This option, which Kripke (1980) himself seemed to defend, suggests that *all* our imaginings represent possibilities and that with respect to the Kripke-Putnam cases our imagination is *mistaken*. That is, we *wrongly* think that we imagine a Kripke-Putnam *a posteriori* impossibility, whereas in reality we are imagining a closely related *possibility*. This suggests an *error-theory* for imagination.

I will argue that we should *not* accept such an error-theory with regards to our imagination. These arguments are unlikely to convince a hardened error-theorist, however, all that I need is that they make the alternatives plausible; motivating the discussion following in the imagination-part of the dissertation.

2.4 Error-Theories of Imagination

Kripke (1980) himself adopted an error-theory with regards to imagination (see Kung, 2016, sec. 1 for a discussion). The idea is that every time you think you have imagined an impossibility, you are mistaken about what you think you're imagining: you actually imagined something that is possible, but (almost) indistinguishable from the impossibility that you think you are imagining (Hill, 1997; Kung, 2016).[19] Kripke, on a imagining the impossibility of a wooden lectern being made out of ice, points out that

> one could have the illusion of contingency in thinking that this table might have been made of ice. We might think one could imagine it, but if we try, we can see on reflection that what we are really imagining is just there being another lectern in this very position here which was in fact made of ice. (1971, p. 157)

Gendler & Hawthorne (2002b, pp. 33-38) discuss these *illusions of possibility* very clearly and note two potential sources of such 'mistakes'. First of all, there might be, what they call, *reference-fixing surrogates*. If we take the example of 'Hesperus is Phosphorus' and the intuition that we seem to be able to imagine that Hesperus is *distinct* from Phosphorus, the reference-fixing surrogate explanation suggests that "our modal intuitions will go astray in so far as we conflate a reference-fixer with the term it introduces" (Gendler & Hawthorne, 2002b, p. 34). That is, we mistakenly think that the heavenly body that occurs in the morning sky might be distinct from the heavenly body that occurs in the evening sky, "[b]ut that contingent truth shouldn't be identified with the statement that Hesperus is Phosphorus" (Kripke,

[19] According to Kung (2016), part of the motivation for this view is what he calls a *telescopic* view of imagination: imagination is a lens through which we look at possibilities. On this view, it is obvious that we cannot imagine impossibilities. What is interesting is that, *if* this is indeed what motivated Kripke, it seems to go against Kripke's *stipulative* view of what possible worlds are (see Berto & Schoonen, 2018, sec. 5 & 6).

1980, p. 105). We mistakenly consider the reference-fixer (e.g., 'the heavenly body in the morning sky') with the terms it introduces (e.g., 'Phosphorus'). However, this strategy of explaining our mistakes concerning what we think we imagine relies heavily on the idea that these terms are introduced by descriptive reference-fixers, whereas Kripke does not think that all (rigid) terms are so introduced; many of them are introduced through baptism by ostension (Kripke, 1980).

The most discussed explanation of our modal errors concerns *qualitative indistinguishability* or, as Gendler and Hawthorne call it, *epistemic duplicates.* The idea is very simple: whenever you think that you are imagining an impossibility, you *actually* are imagining a possible situation that is (qualitatively) indistinguishable from the impossibility you think you are imagining. To take the Hesperus and Phosphorus example again:

> There certainly is a possible world in which a man should have seen a certain star at a certain position in the evening and call it 'Hesperus' and a certain star in the morning and call it 'Phosphorus'; and should have concluded—should have found out by empirical investigation—that he names two different stars, or two different heavenly bodies. [...] And so it's true that given the evidence someone has antecedent to his empirical investigation, he can be placed in a sense in *exactly the same situation, that is a qualitatively identical* epistemic situation, and call two heavenly bodies 'Hesperus' and 'Phosphorus', without their being identical. (Kripke, 1980, pp. 103-104, emphasis added)

We *think* that we have imagined a world where Hesperus is not Phosphorus, but all we've imagined is a qualitatively similar world where a planet that appears in the evening and is called 'Hesperus' is distinct from another planet that appears in the morning sky and that is called 'Phosphorus'. We didn't *really* imagine Hesperus or Phosphorus.

In general, error-theorists hold that we are mistaken about what we think we imagine when we imagine impossibilities and that, in general, our unrestricted imagination is a reliable guide to possibilities, we just fail to appreciate this from time to time.

2.4.1 Against Error-Theories of Imagination

It is important to understand that an error-theory about imagination is a universal claim: *each* time you think you imagine an impossibility, you are mistaken in what you imagine. So, the claim is about imagination *irrespective* of its role in the epistemology of modality. It is not that error-theorists hold that you can imagine impossibilities in general, but when we engage with the epistemology of modality, it turns out that we are mistaken about what we imagine when imagining impossibilities. The error-theorist thus has to explain away *all* our intuitive imaginings about impossibilities as mistaken. This means that we can evaluate the error-theory independently of our modal epistemological intuitions. As Kung puts it, setting our modal epistemology aside for the moment, "[h]ow plausible is it that we cannot imagine certain impossible situations, and that we make mistakes about what we imagine when we try to?" (2016, p. 94).

I agree with many that this is highly implausible, we (seem to) imagine impossibilities with ease (see, e.g., Hill, 1997; Wright, 2002; Byrne, 2007; Fiocco, 2007; Kung, 2016; Priest, 2016; Berto & Schoonen, 2018; Wright, 2018). Consider the following example. You have to pick up a guest speaker at the airport and all you know is that their name is 'Quinn'. You stand there imagining that Quinn is a blonde man, but when they arrive, it turns out that you were wrong and Quinn is a woman. When you meet her, "[y]ou might laugh and tell her, 'I imagined that *you* were a man!"' (Kung, 2016, p. 95, original emphasis). If biological sex is a Kripkean *a posteriori* necessity, which many take it to be, then, according to the error-theorist, this is wrong; you did not imagine *her*, you imagined someone that you mistook for Quinn. But this does not seems right; there is no doubt in your mind that you imagined her, Quinn, and not some other person. (See Priest, 2016, p. 195 for further examples.)

The trouble with error-theories of imagination is that the claim that we might be mistaken about what we imagine is in tension with the idea that imagination is under our voluntary control (Kung, 2016; Langland-Hassan, 2016; Williamson, 2016a). "As a general rule I get to say who my imagination is about" (Kung, 2016, p. 103, fn. 27). If I get to say who or what my imagination is about, how can it then be that I am mistaken about what I imagine? Consider what Wittgenstein says what we do when we encounter someone who says they've imagined King's College being on fire.

> We ask him: 'How do you know that it's King's College you imagine on fire? Couldn't it be a different building, very much like it? In fact, is your imagination so absolutely exact that there might not be a dozen buildings whose representation your image could be?' (1958, p. 39)

However, in addressing this issue, surely we should not doubt our imagination. Wittgenstein continues, "[a]nd still you say: 'There's no doubt I imagine King's College and no other building' " (ibid.).[20] This idea that we are perfectly aware of what our imagination is about is echoed through the literature. For example, "Kripke's explanation [...] is fundamentally misguided; for as I see it, in non-pathological

[20] Interestingly, the certainty about what it is that you imagine has the same source as Wittgenstein's pessimism about the epistemological value of imagination: imagination is under our voluntary control. Because we choose what it is that we imagine, we are not wrong about it. But it is also for this reason, Wittgenstein thought, that imagination cannot provide us with justification: "It is just because forming images is a voluntary activity that it does not instruct us about the external world" (Wittgenstein, 1967, §627).

circumstances introspection gives us pretty accurate access to the contents of our own states of imagination" (Hill, 1997, p. 83, fn. 10)" and "[a]s a general rule, when we imagine something there is just no doubting that we have imagined that something" (Kung, 2016, p. 95).

Wright (2002, 2018) presents another example that seems especially hard to explain away for error-theorists. Let's assume, with Kripke, that biological origins are essential. Wright argues that I can nevertheless imagine myself as having been born from different parents. I can even imagine myself "originating in a different world, of a different race, and having been visited on Earth from afar and brought up as their own by the people whom I take to be my biological father and mother" (Wright, 2002, p. 435). These imaginings, Wright argues, do not allow for the error-theorists' explanation, as "[n]o mode of presentation of the self need feature in the exercise before it can count as presenting a scenario in which *I*" have those origins (Wright, 2002, p. 436, original emphasis).[21] The general lesson is that there seem to be imaginable impossibilities that rely on a first-person perspective where we cannot "fail to be sensitive to the distinction between [ourselves] and a mere epistemic counterpart, a mere 'fool's' self, as it were" (Wright, 2018, p. 271).

I take it that these examples show that, modal metaphysics and epistemology aside, it is very intuitive that we can imagine impossibilities and that it is highly unlikely that in all these cases, we are mistaken about what we think we've imagined. Casting error-theories of imagination aside while maintaining imagination as a basis for our epistemology of possibility means that we have to find some way of ruling out imagining the Kripke-Putnam impossibilities, lest our imagination-based epistemology of possibility suggest that we can justifiably believe these impossibilities to be possible. The first chapter of imagination-part of the dissertation focuses on a theory of imagination that explicitly aims to do just that. I argue that they fail unless they rely on problematic prior modal knowledge. Consequently, we will turn to discuss restrictions through recreativist imagination in Chapter 4 and Chapter 5.[22]

[21] Some versions of Wright's example are somewhat controversial. For example, the version prompting the strongest intuition involves imaginings of completely different selves (e.g., being a monkey, see Berto & Schoonen, 2018), yet, it is not obvious that we would in fact be imagining that we *are* monkeys, rather than that we are humans pretending to the best of our abilities to be(have like) monkeys (see Nagel, 1974). However, examples of imagining originating from different human parents might be less problematic and already enough.

[22] Where, as we will see, the issues of Kripke-Putnam cases do not play as significant a role.

Chapter 3

The Problem of Modally Bad Company

[S]tipulation has no legitimate role to play [in the question of] how we know what is modally true

– Divers, 2002

In this chapter, I will discuss theories of imagination that focus on the *representational content* of the imaginings, which have a long tradition as imagination-based epistemologies of possibility.[1] I will focus on a particular family of theories of imagination that (i) use linguistic content to distinguish between qualitatively indistinguishable imaginings and (ii) provide a basis for a significant epistemology of possibility that give the right predictions on Kripke-Putnam cases *without* reliance on an error-theory. They aim to do so by *restricting* the linguistic content allowed in imaginings that feature in their epistemology of possibility.

I will argue that even these sophisticated accounts of imagination *fail* to provide a satisfactory basis for an epistemology of possibility. In particular, I will argue that there is a deep methodological problem that these accounts face: in order to deliver the significant epistemology of possibility that they promise, they have to rely on problematic prior modal knowledge. This leads the way to investigate radically different conceptions of imagination in the next two chapters.

[1] The material of this Chapter is based on Schoonen (2020).

3.1 QALC Imagination

An intuitive view of imagination is that it *represents* (hypothetical) situations (Yablo, 1993; Chalmers, 2002; Kung, 2010; Dohrn, 2019). We will call this the *Representational View of Imagination.*[2] For example, when you imagine yourself playing basketball, your imagination represents to you a situation where this is so. Note that we want to make sure that if you imagine yourself playing basketball, it is *you* whom you imagine and not some qualitative duplicate. That is, we want a theory of imagination that "overcom[es] some of the shortcomings" of an "image-based account of imagination" (Kung, 2017, p. 136). In particular, these theories go beyond the purely qualitative content in order to capture Quantity and Aboutness via Linguistic Content.[3] I will therefore call these accounts theories of *QALC imagination.*

Importantly, if we take imagination, on such a view, to be the basis for our epistemology of possibility, we have to be able to distinguish between the imaginings that represent *possible* situations from those that represent *impossible* situations. In particular, we should want to be able to rule out those imaginings that represent Kripke-Putnam *a posteriori* impossibilities (see Chapter 2, Section 2.3). To that end, QALC imagination theorists only allow imaginings with *restricted* linguistic content to play a role in their epistemology of possibility; namely, linguistic content that is "*grounded in the right way in actual experience*" (Kung, 2017, p. 136, emphasis added).[4]

This short, intuitive discussion of the representational view of imagination and epistemologies of possibility based thereon gives us the two desiderata for a QALC imagination-based epistemology of possibility. The first one being that the theory of imagination goes beyond the limitations of a Humean, imagistic account of imagination, where imagination *only* has qualitative content without any linguistic content. On such a Humean account, one can no longer distinguish Wittgenstein himself from a qualitative duplicate of him; or distinguish the imagining of two mono-zygotic twins Quinn and Blake, where Quinn sits next to a standing Blake, from one where Blake sits next to a standing Quinn. QALC imagination theorists aim to improve upon such a picture by capturing numerical distinctness via *linguistic content.*[5]

[2] This is not to say that the kinds of imagination discussed in Chapter 4 and Chapter 5 *deny* that there is representational content to imagination, it is just that this is not the *focus* of these account. This is merely a distinguishing label; nothing theoretically significant should be derived from it.

[3] There are multiple phrases used to denote this kind of content, e.g., 'assigned content', 'stipulated content', et cetera; I will use 'linguistic content' as I feel it is the least misleading and I intend to remain non-committal about what it is exactly.

[4] Often, when I talk of 'QALC imagination theorists', I mean 'theorists who provide a QALC imagination-based epistemology of possibility'.

[5] In Appendix A, I discuss such an 'image-based' account of an imagination-based epistemology

Another desiderata of a QALC imagination theory is that they give the correct predictions for Kripke-Putnam cases (i.e., *a posteriori* impossibilities) without appealing to an error-theory. "[I]maginability is a guide to possibility," Byrne (2007) argues, "only if Kripkean impossibilities are unimaginable" (p. 130). Similarly, Kung points out that "[a] virtue of this account is that it dovetails with the Kripke-Putnam thesis about *a posteriori* necessities" (2010, p. 650). And Gregory says that "[p]utting together the above remarks on the plausible instances of simple a posteriori refutable impossibilities, we get that nothing will plausibly be viewed as a simple a posteriori refutable impossibility which is unshakeably imaginable" (2004, p. 335).

We use these two desiderata to formulate more precisely what we take QALC imagination to be.

A theory is a *QALC* imagination-based epistemology of possibility if

1. It distinguishes between qualitative indistinguishable imaginings via linguistic content.[6]

2. It aims to give the correct predictions on Kripke-Putnam cases without appeal to an error-theory.

These two criteria capture exactly the two points at which QALC imagination theorists aim to improve upon traditional theories of imagination and imagination-based epistemologies of modality. Traditional image-based accounts cannot distinguish between qualitatively indistinguishable imaginings, hence criterion **1** and traditional imagination-based epistemologies give the wrong predictions on the Kripke-Putnam cases, hence criterion **2**.[7]

When defined in such a way, many theories fall under the QALC imagination label. For example, early epistemologists of modality who tried to define what exactly it is to *conceive* of something. Such as Yablo (1993) – "p is conceivable for me if I can imagine a world that I take to verify p" (p. 29) – and Chalmers (2002) – "[o]ne modally imagines that P if one modally imagines a world that verifies P, or a situation that verifies P" (p. 151) (but also Van Cleve, 1983; Tidman, 1994; and Hill, 1997).[8] More recently, people have started to develop this kind of

of possibility (see also Kung (2017, especially sec. 8.4) for a detailed description of the Humean account and the improvements thereupon by accounts of QALC imagination). In that appendix, I also discuss the issues that arise when one allows for *unrestricted* linguistic content. Those who present a QALC imagination-based epistemology of possibility aim to walk a middle ground between these two.

[6] Equivalently, it captures numerical distinctness ('aboutness') through linguistic content.

[7] See Yablo (1993) for a discussion of a variety of attempts of dealing with the Kripke-Putnam cases.

[8] Though remember from Chapter 2 (Section 2.1.1) that one might also think that Chalmers

imagination *independently* from the idea that it is the correct way of spelling out what conceivability is. Examples of such theorists are Geirsson (2005) – "[w]hat is important is that regardless of whether one uses propositional or pictorial imaging one can construct *scenarios*" (p. 293, original emphasis) – and Dohrn (2019) – "[o]ne is justified to believe that p is possible if one entertains a suitably concrete and consistent representation of a world which one takes to verify p" (p. 8). To give a sense of how widespread the idea is, the following authors all discuss (and sometimes defend) theories that, according to the above definition, are theories of QALC imagination: Kripke (1980); Gregory (2004); Byrne (2007); Fiocco (2007); Stoljar (2007); Doggett & Stoljar (2010); Gregory (2010); Kung (2010); Hartl (2016); Lam (2017).[9]

QALC imagination represents a situation using qualitative and linguistic content. Qualitative content is similar to that of perception in that it presents the "'basic observational' properties in imagined space" (Kung, 2010, p. 624). Some basic qualitative properties are shape, colour, distribution in space.[10] This allows us to account for imaginings in which there is a qualitative thing that looks exactly like Wittgenstein and is blonde. However, remember that this does not yet allow us to imagine that it is *Wittgenstein* who is blonde as opposed to a mere qualitative duplicate of him. In order to account for this, imaginations are considered to have a second component of content: *linguistic content*.

The linguistic content does a lot of work in these theories, so let us discuss it in a bit more detail. Linguistic content is, roughly, content that comes with qualitative content. Kung (2010), for example, distinguishes between different types of linguistic content: *labels* and *stipulations*. Labels are very simply (linguistic) labels that 'attach' to the things in the qualitative imagining. So, when I imagine Susan giving a lecture, I do not only imagine a thing qualitatively similar to Susan, I am sure that I am imagining *her*, Susan. This is secured through the label, 'Susan', that accompanies the qualitative content. Stipulations, on the other hand, are propositional contents that go "above and beyond that of the mental image" (Kung, 2010, p. 625), i.e., that do not 'attach' to specific parts of the qualitative content. When, for example, I imagine Andy and Susan meeting as friends, their friendship, that they meet on a Friday, and that they speak English are all stipulated content.[11]

holds conceiving to be a significantly distinct cognitive phenomenon, depending on one's view of what imagination is. In that case, it is not clear whether Chalmers' view can be classified as a QALC imagination.

[9] Some of these authors reject the idea that QALC imagination is a guide to the possible (e.g., Byrne, 2007 and Fiocco, 2007) and other suggest that it is QALC imagination *in addition* to something else (e.g., Hartl, 2016 and Dohrn, 2019). Yet all of these authors do discuss QALC imagination.

[10] Let's assume for the sake of the argument that *sortal properties* can be part of the qualitative content (or at least up to a certain point). So, that I imagine a purple *cow*, as opposed to a purple cow-shaped object, is part of the qualitative content (Siegel, 2006).

[11] Thanks to an anonymous referee of (Schoonen, 2020) for encouraging me to make the difference

(Note that this distinction between labels and stipulation is not essential for a QALC account of imagination.)

There are two things that deserve some emphasis. First of all, it is very important to stress that even though we can pull apart these two kinds of content – i.e., qualitative and linguistic – in analysing imagination, cognitively speaking imagination is *not* a two-stage process. We do not imagine a qualitative scene and then add the linguistic content. We imagine a situation with all its content in one go – i.e. the "imagery comes with everything already labelled and stipulated" (Kung, 2010, p. 625).

Secondly, the above does not presuppose, what Wiltsher (2016) calls, *the additive view* of imagination. That is, even if one thinks that imaginings do not have two distinct content components, they could still, more or less, fall under this description of QALC imagination. For example, Wiltsher (2016) argues against the two content components in imagination (he argues that there is only qualitative content), but even he still accepts that there are sometimes imaginings that have linguistic content. Conversely, Hutto (2015) suggests that there is only linguistic content in imaginings, but even he agrees that these can be accompanied by instances of mental imagery.[12] What is crucial for our purposes is that there can be *some* linguistic content. To see what I mean, consider the discussion of Kung's labels by Wiltsher (2016). He argues that much of the labels of Kung's theory are already present in the qualitative content. That is, Wiltsher suggests a very rich kind of qualitative content. Yet, on his account, we still need to distinguish between imagining Susan and Andy and imagining Susan and Andy *being second cousins*. Similarly, we still need to be able to distinguish imaginings of Mark Twain hitting Samuel Clemens and imaginings of Mark Twain hitting some qualitative duplicate of Samuel Clemens. I take it that on Wiltsher's account these imaginings are cases where his "view need not deny that non-sensory imagining can accompany mental images, and that the two together can provide a richer imaginative experience" (2016, p. 275).

3.2 QALC Imagination and Epistemology of Possibility

QALC imagination theorists hold that imagination can justify our beliefs in what is possible, while acknowledging the Kripkean *a posteriori* necessities and rejecting an error-theory for imagination. This means that they have to address two questions: (i) can their theory account for our intuition that we can imagine impossibilities? and (ii) how does this account play a role in the epistemology of possibility? Note that on the face of it, these two questions seem in immediate tension with each

between labels and stipulations clearer.

[12] For a nice discussion of both these arguments and their relation to the additive view, see Tooming (2018).

other. We will start by discussing the answer to the first question, which will then give rise to the second.

How we are able to imagine the impossible is easily explained on these accounts: by means of the linguistic content. There are plenty of imaginings that represent impossibilities and in most cases it is the linguistic content that is doing the work. Consider for example the following imaginings:

(1) Frank Zappa is my father.

(2) Mark Twain is fighting Samuel Clemens.

(3) Susan is a cleverly disguised robot.

The linguistic contents involved are nothing out of the ordinary. Imaginings that combine these linguistic contents with particular qualitative content, are imaginings of impossibilities (e.g., Priest, 2016; Berto & Schoonen, 2018; and Appendix A.2).[13]

So, how is it that QALC imagination can play a role in the epistemology of possibility if we can so easily imagine impossibilities?

3.2.1 Authenticating Linguistic Content

Remember that in order for imagination to be epistemically useful, it needs to be restricted (e.g., Kind, 2016a; Kind & Kung, 2016a; Williamson, 2016a; Balcerak Jackson, 2018). Though QALC imagination theorists are seldom explicit about this (with Kung, 2010, 2016, 2017 a notable exception), for them restricting imagination is allowing only *certain kinds* of linguistic contents in imaginings that are supposed to justify our beliefs about what is possible.[14] Only linguistic content for which we have *independent* evidence that it is possible is allowed in imaginings that play a role in our epistemology of possibility. That is, we need to verify or authenticate the relevant linguistic content. The verification happens *recursively* – i.e., there can be verification through imagination as well as verification through "some other source" (Kung, 2010, p. 642). The main source of verification that is appealed to is *evidence from actuality* (e.g., Gregory, 2004, 2010; Kung, 2010; Dohrn, 2019). As an example, this is how one might apply this method to show that we are justified in believing that Andy and Susan could be distinct, based on imagining it:

[13] Whether or not qualitative content *alone* can represent impossibilities is open for discussion. Some think that the Escher-like paintings are a prime example of a purely qualitative impossibility (Kung, 2010; Balcerak Jackson, 2018), yet others suggest that this is just a collection of possible qualitative contents that are jointly inconsistent (Sorensen, 2002). We need not engage in this discussion, it is fine for our purposes if it is only "*by dint of assignment* that we are able to imagine an impossible situation" (Kung, 2010, p. 636, emphasis added).

[14] As Kung (2010) is one of the few how explicitly discusses this part of QALC imagination theories and does so in a very elaborate fashion, I will focus mostly on his discussion here, yet the problems raised below apply to all QALC imagination-based epistemologies of possibility.

> One needs to authenticate that Andy could possibly exist and does so by appealing to the actual existence of Andy. This allows us to use the label 'Andy'.[15] The same goes for Susan. Further, one can "appeal to the actual diversity [of Andy and Susan] to satisfy [the distinctness] demand" (Kung, 2010, p. 644).[16]

The recursive authentication seems to work quite well. Our imagination that Susan and Andy are distinct can justify our belief that they could be distinct. However, as we cannot authenticate the distinctness of water and H_2O, QALC imagination does *not* justify us to believe that it is possible that water is not H_2O. This means that QALC imagination seems to be able to accommodate **2**: they seem to be able to deal with Kripkean *a posteriori* impossibilities. This is the hallmark of QALC imagination-based epistemologies of possibility (remember the quotes from Gregory, 2004; Byrne, 2007; and Kung, 2010 discussed above).

3.3 The Problem of Modally Bad Company

I will argue all is *not* well and that there is a deep methodological problem with these QALC imagination-based epistemologies of possibilities. The problem lies at the core of these accounts, as it concerns the combination of (authenticated) linguistic content and qualitative content. Different kinds of examples could highlight the problem, but it is best expressed by considering a pair of imaginings: one representing a mundane possibility and one representing an *a posteriori* impossibility. I will call such imaginings pairs of *modally bad company* and the resulting problem, the *problem of modally bad company*.[17] In a nutshell, the problem shows that QALC imagination-based epistemologies of possibility cannot allow linguistic content, even when it is authenticated, without reliance on problematic prior modal knowledge. Given the

[15] Note that it is not trivial what the right account of 'labelling' is. Remember that the image comes "with everything already labelled" and that we are ourselves in charge which labels accompany the image – i.e., that imagination is up to us (Kung, 2010, 2016). I take it that a very natural understanding of labels is that we are certain to which parts of the qualitative image the labels apply.

[16] It might seem strange that distinctness is treated as linguistic content rather than qualitative content. Here is why. One need not authenticate that there is a qualitative occupant in space, this is the qualitative content. However, "[w]hat needs authentication is the *identity* of the thing," i.e. to what object it relates (Kung, 2010, p. 643, original emphasis). In line with criterion **1**, "*identities* are non-pictorial [i.e., linguistic] content. [. . .] [T]he image doesn't, in virtue of its qualitative features, depict particularity. The image does not distinguish between qualitatively identical tokens of the same type" (Kung, 2017, p. 146, original emphasis).

[17] The problem is, in a way, inspired by the Bealeresque (Bealer, 2002) comments made by Cameron (2010) in a response to Gregory's (2010) work. The label for the problem is inspired by the *unrelated* problem of bad company for Neo-Fregeans (see Linnebo, 2009 and Tennant, 2017, fn. 19). Thanks to Franz Berto and Thomas Schindler for suggesting this label to me and pointing me to the relevant literature respectively.

way that I will raise the issue here, with pairs of modally bad company imaginings, the problem presents itself as a dilemma: either QALC imagination theorists *fail* to justify a wide range of mundane possibility statements, resulting in (a form of) *radical modal scepticism*, or QALC imagination theorists have to rely on problematic prior modal knowledge.[18] In its most general form, the problem suggests that there is a tension between the two core criteria of QALC imagination theories: by allowing linguistic content, one cannot rule out the specific Kripke-Putnam cases without reliance on prior modal knowledge.

The first half of a modally bad company pair concerns an *a posteriori*, non-actual, mundane possibility claim that requires linguistic content to be imagined. These kinds of cases are significant in number – involving, e.g., distinctness claims concerning non-actual possibility (Mark Twain is distinct from his non-actual twin); constitutional claims concerning non-actual possibilia (my non-actual pet dog being a dog); non-actual constitutional claims about actualia (my non-actual metal hip); non-actual mental states of actualia (my non-actual headache); et cetera. Consider the following imagining as an example of such a case:

Modally Innocent

(4) Imagine that Mark Twain is playing basketball with his, non-actual, twin brother: Mark Twin. Mark Twain is jumping higher than his brother.

This is a mundane possibility claim: someone having a sibling more than they actually have (or, even more generally, the possibility of non-actual things). I take it that any epistemology of possibility ought to predict that the beliefs in such ordinary, mundane possibility beliefs are justified. Collectively, these cases constitute a large part of the class of mundane possibility claims such that "a theory of modal epistemology or modal metaphysics is likely to be viewed with suspicion if it suggests that we are *not justified in believing [them]*" (Hawke, 2011, p. 360, emphasis added). I contend that if QALC imagination theories *fail* to account for these kinds of cases, their appeal as a promising epistemology of possibility is severely undermined; irrespective of whether they manage to get the right predictions concerning the Kripke-Putnam cases.

Luckily, QALC imagination theories have the tools to authenticate the linguistic content such that these imaginings count as evidence for their possibility. As authentication by actuality is not possible here (except perhaps for the label 'Mark Twain'), the recursive procedure needs to appeal to something else. Kung puts the

[18] Maybe a weaker worry would already be problematic enough: there is an epistemic asymmetry between the members of a modally bad company pair that the QALC imagination theories cannot capture. This concerns the mundane and controversial nature of the two cases involved. However, as it is not obvious that the QALC imagination theorists are concerned with the different modal status of imagined situations (Van Inwagen, 1998; Hawke, 2011), I stick with this stronger worry.

issue as follows, “how do we authenticate the assignment *something is X* when X does not exist? [. . .] The only option is to imagine a situation lacking that assignment where it is intuitive that one of the imagined things is X” (2010, p. 652). One way to do so is to imagine a generic, but obviously possible (i.e., with authenticated linguistic content) story about how two individuals would be distinct. Imagining diversity of origins would do the trick. So, we imagine Jane Clemens conceiving Mark Twain *and* Mark Twin; Jane’s being pregnant; her giving birth; and her and John Clemens holding the twins. The labels of Jane and John Clemens can be authenticated by appeal to actuality and I take it that we can imagine the ‘baptism by ostension’ (Kripke, 1980) of the labels for Mark Twain and Mark Twin.[19] Let us call this explanation **Conception**.

It seems that if the QALC imagination theorists want to account for our justified beliefs in ordinary possibility claims such as (4), they have a plausible story to tell. The problem of modally bad company is that the same story seems to be able to justify our belief in the modally bad counterpart of (4).

Consider the modally bad counterpart of (4):

Modally Suspicious

(5) Imagine that Mark Twain is playing basketball with Samuel Clemens. Mark Twain is jumping higher than Samuel Clemens.

At best, we should take our epistemology of possibility to be agnostic on modal status of (5) (Roca-Royes, 2017); at worst (5) is impossible (as many QALC imagination theorists seem to think). Either way, our epistemology of possibility should *not* judge imagining (5) to provide us with evidence for its possibility. However, as I mentioned, the problem of modally bad company suggests that something like **Conception** seems to allow us to move from imagining (5) to justifiably believing its possibility, *unless* we rely on problematic prior modal knowledge.

To see this, let us see how a story similar to **Conception** applies to (5). We assume that the qualitative content in (4) and (5) is insignificantly different and that the relevant linguistic content is explicitly mentioned in the case description. What is crucial is the numerical distinctness between Mark Twain and Mark Twin/Samuel Clemens, which, as you remember, is what motivates core feature **1** of QALC imagination theories.[20] In **Conception**, we recursively imagined distinct origins in order to justify the distinctness of Mark Twain and Mark Twin. For (5) we can do the

[19] One has to tell a story about the labels for actually non-existent objects and this seems as good as any. See Kung (2010, p. 653) for a similar story on authenticating labels for non-existent objects.

[20] Note, it seems that we can authenticate the label ‘Samuel Clemens’ by appeal to actuality, whereas we can not authenticate ‘Mark Twin’ this way. This does seem to be a way to distinguish between the two cases, but this suggests that (5) is verifiable and (4) is not, which does not line up with our pre-theoretic intuitions.

same: we imagine the distinctness of the origins (by way of recursive imagination) of Mark Twain and Samuel Clemens and then we either appeal to actuality for the labels or imagine the baptism.[21]

It seems as if QALC imagination theories justify our belief that (5) is possible in the same way it justified our belief in the possibility of (4). However, (5) is a paradigm instance of an *a posteriori* Kripke-Putnam impossibility – i.e., the kind of case that QALC imagination theories are committed to getting right.

One might think that the issue is just an idiosyncrasy of this particular example. This is not the case. In fact, we can construct a whole range of modally bad company pairs and "it is difficult to assess how widespread the problem [. . .] really is" (Balcerak Jackson, 2018, p. 215).

- *A posteriori distinctness claims between actual objects and non-actual possibilia*:

  Modally Innocent: Some actual x is distinct from some non-actual y.

  Modally Suspicious: Mark Twain is distinct from Samuel Clemens.

- *Non-actual mental states of agents*:

  Modally Innocent: Some actual y has a non-actual mental state Φ (e.g., I could have a headache even though I actually don't).

  Modally Suspicious: Mark Twain is a philosophical zombie.[22]

- *Constitutional claims about non-actual possibilia*:[23]

  Modally Innocent: Some x having a non-actual prosthetic P (e.g., my non-actual metal hip).

  Modally Suspicious: Mark Twain is a cleverly disguised robot.

In all these cases, the general problem, which gets at the core of QALC imagination-based epistemologies, comes to light. For non-qualitative, *a posteriori* non-actual situations, we need to combine qualitative content with linguistic content (often where the former justifies the latter); yet there is no principled way to *rule out*

[21] "One might protest, in Kripkean fashion, that the [people] wouldn't be [*Mark Twain* and *Samuel Clemens*]. But in my imagining, I am not leaving it open whether or not [they are]. As a general rule I get to say who my imaginings are about" (Kung, 2016, p. 103, fn. 27).

[22] I am *not* claiming that philosophical zombies are *a posteriori* impossible, just that it is impossible for Mark Twain to be one.

[23] If you think, contra Siegel (2006), that sortal properties are *not* part of the qualitative content, then we can extend this problem even further. In that case, one needs to authenticate that my non-actual pet dog could be a dog. However, we can run the problem of modally bad company and use the same methods that we use to authenticate this to authenticate that Mark Twain is a dog. (Note that appeal to the perennialness of this property – i.e., the fact that once acquired, it is never lost – does not help, for not all perennial properties are necessary: e.g., being dead.)

certain combinations of qualitative and linguistic content while allowing others (i.e., their modally innocent counterparts).[24]

This raises a dilemma for QALC imagination-based epistemologies of possibility:

Sceptical-horn: Reject the explanation of authentication for the modally innocent cases.

Acceptance-horn: Accept the explanation of authentication for the modally innocent cases.

The sceptical horn results, as we saw above, in unwarranted radical modal scepticism for a significant part of ordinary possibility claims (e.g., I could have a headache). This undermines the theory as a serious epistemology of possibility. So, QALC imagination-based epistemologies of possibility should opt for the second horn: accept that their theory allows the modally innocent cases to justify the resulting modal beliefs. This leaves them with two options with regards to the modally suspicious cases. (i) They accept that these are *also* evidence for their possibility, but try to explain this away. (ii) They try to come up with a distinguishing feature that allows them to differentiate between the two cases; this, I argue, can only be done by reliance on prior modal knowledge.

I will discuss these options in turn in the next two sections.

3.4 The Fallibilism Response

Could the QALC imagination theorists accept these findings without too much trouble? That is, can they accept that in the modally suspicious cases, their theory gives the wrong predictions about whether we should be justified in believing their possibility?[25] Most QALC imagination theorists take their theory to be *fallible*: based on their theory's prediction, we are allowed to justifiably believe "that p even though one's evidence does not guarantee the truth of p" (Brown, 2018, p. 2) (see Leite, 2010). So, one might argue, all you have shown us is what we already acknowledge, "there are cases where imagining even according to [our theory] will lead to an incorrect judgment about possibility" (Kung, 2010, p. 658).

As we take the imagination involved in QALC theories to be our ordinary capacity to imagine things, it is clear and, in itself, unproblematic that the resulting

[24] The issue can also be brought to light with other examples. I have a qualitative imagining of a Mohammad Ali-like object punching a Cassius Clay-like object. Why is it that I am allowed infer possibility of an instance where I label the two objects 'Mohammad Ali' and 'Cassius Schmlay', but *not* when I label them 'Mohammad Ali' and 'Cassius Clay'?

[25] Thanks for an anonymous of (Schoonen, 2020) reviewer for pressing this response on behalf of the QALC imagination theorists.

theories present a fallible epistemology of possibility. However, when QALC theorists use fallibilism as a response to the problem of modally bad company, it cuts out the heart of their own theory. Remember, the modally suspicious cases are instances of *a posteriori* impossibilities; negations of Kripke-Putnam cases. These were exactly the kinds of cases that *motivated* the QALC imagination theory as superior to a naïve, purely qualitative imagination-based account (Kung, 2017). In particular, if their theory fails to give the correct predictions with respect to these Kripke-Putnam cases, they fail to satisfy the crucial criterion **2**. According to Byrne – who said that " 'imaginability is a guide to possibility' only if Kripkean impossibilities are unimaginable" (2007, p. 130) – these theories would no longer be proper epistemologies of possibility. Of course, one may suggest that in these particular cases some *other source* of modal knowledge should overrule our evidence from imagination, but given the sheer number of these cases, this significantly undermines the attempt to "explain how a very reasonable epistemology of possibility *flows from* a theory of imagination" (Kung, 2010, p. 621, emphasis added).[26]

3.5 The Differentiating Response

In this section, I will discuss a number of methods that a QALC imagination theorist might appeal to in order to *distinguish* between the two imaginings of the modally bad company pair. I argue any *successful* method relies on problematic prior modal knowledge.

Follows from Linguistic Content alone

The QALC imagination theorists might suggest that if something follows *from the linguistic content alone*, then we should not be justified in believing that that thing is possible (what, e.g., Kung means with 'from the assignment alone' is that the conditional '[linguistic content] → [imagined proposition]' would be true, see his p. 640). However, it is unclear why, if it works, this condition would rule out the problematic case and *not* the good case. In both cases the same (kind of) claims follow from the linguistic content alone. If we are supposed to rule out the modally suspicious case on this basis, we should rule out the innocent case on the same grounds. Hence, this method fails to discriminate between the two cases.[27]

[26] Additionally, the cases that QALC imagination theorists would be sweeping under the fallibilist rug do not seem to be the kind of cases that they have in mind when they themselves suggests their theories to be fallible. They have in mind cases of where the *qualitative* content is misleading evidence, for example in cases of the famous Escher drawings (see for example Kung, 2010, p. 658).

Moreover, the sheer *number* of *a posteriori* impossibilities that we are seemingly able to justifiably believe to be possible is too numerous to sweep under the fallibilism rug (even if one takes the more statistical interpretation of fallibilism, e.g., Lam, 2017).

[27] Additionally, it is not obvious that 'following from linguistic content alone' would rule out (5) as evidence for its possibility. For one, the content does not follow from the linguistic content

Absolute Certainty

One might suggest that if one is *absolutely certain* about something, then if we imagine its negation this should not justify our believing the possibility thereof. This might be a *prima facie* plausible additional condition: if I am absolutely certain that $2 + 2 = 4$, then even imagining it otherwise should not justify me in believing the possibility of $2 + 2 \neq 4$. However, the notion of absolute certainty is extremely strong. For example, Kung (2010, p. 629) mentions that I should not even be absolutely certain that I am Tom. It thus seems unlikely that we are absolutely certain that Mark Twain does not jump higher than Samuel Clemens.

Conceptual (In)Coherence

Another sensible additional condition might be that if there is conceptual incoherence in an imagined scenario, that imagining should not justify any beliefs in what is possible. For example, imagining that there is a maple-leaf shaped oval does not justify one in believing that maple leaf-shaped oval are possible on the basis of conceptual incoherence (Yablo, 2002; Weatherson, 2004). However, there is clearly no conceptual incoherence in the thought that Mark Twain jumps higher than Samuel Clemens.[28]

Unwillingness

Some have suggested that we might be unwilling to imagine certain things and that this is something that we need to take into account (Gendler, 2000; Weatherson, 2004). Again, this is clearly not the case with these scenarios.

Appeal to Actuality

The reason why (5) is impossible, one might suggest, is that *in actuality* Mark Twain is Samuel Clemens (see Van Inwagen, 1998, p. 74, fn. 11 for something like this); therefore they *cannot* be distinct. Such an approach hinges, implicitly, on prior *modal* knowledge. To see this, apply this reasoning to (4): because in actuality Mark Twain is twinless, he *cannot* have a twin. Appeal to actuality only works in "joint application [with] the theorem '$x = y \to \Box x = y$' " (Van Inwagen, 1998, p. 74,

alone, we also need the qualitative content (e.g., of the birth of the two people and their jumping). However, one might argue that this is an uncharitable interpretation and that maybe the imaginative content *does* all follow from the linguistic content. There is, I think, an issue with this suggestion. It is unclear why we should think that the imagination should count as evidence for its possibility *at all*, if we recognise that it is only the linguistic content that does the work (see Appendix A.2). Thanks to Pierre Saint-Germier for discussing these issues with me.

[28] Unless one holds that the concept 'Mark Twain' implies something like 'is necessarily identical to Samuel Clemens', but in such a case one use conceptual knowledge to smuggle in knowledge of necessities (see Roca-Royes, 2019a on the issue of modally loaded concepts). Thanks to Deb Marber for raising this issue.

fn. 11); which is knowledge that identities are *necessary*.[29] In order for this method to be successful, we need to know *which properties* are necessary (e.g., identities) and which ones are not (e.g., being twinless), *before* we can judge imaginings to give us evidence for possibility.

Conflicting Modal Intuitions

Finally, then, it might be that we just have a conflicting modal intuition (potentially irrespective of our intuitions about the imagined situation).[30] As with appeal to actuality, this only works, if it does, due to prior knowledge of necessities.

Consider the range of (conflicting) modal intuitions. The conflict does not arise because we find it intuitive that Mark *doesn't* jump higher than Samuel. We also find it intuitive that this chapter does not start with a 'Y' and that Mark Twain doesn't have a twin brother, but this doesn't count as evidence against the *possibility* of the relevant imaginings. If that were so, then we could never gain evidence for *non-actual* possibilities through imagination. For similar reasons, the fact that we may find it intuitive that Mark *possibly* does not jump higher than Samuel is too weak. The only intuition that would 'conflict' is the intuition that Mark *couldn't* jump higher than Samuel: there is no situation, including the imagined one, where Mark jumps higher than Samuel. This modal intuition would indeed defeat the evidence from imagining (5), but requires prior knowledge of a necessity.

3.6 Problem of Prior Modal Knowledge

It thus seems that *if* there are successful methods of discriminating between two cases of a modally bad company pair, it is because of reliance on prior knowledge of necessities. This completely undermines the project of providing an epistemology of possibility, as I will argue in this section.[31]

Remember that QALC imagination theorists aim to provide an epistemology of *possibility*.[32] For any epistemology of possibility, relying on prior knowledge of

[29] "[T]he claim that water is H_2O is metaphysically necessary is supposed to flow from conceptual knowledge that if water is H_2O, it is so necessarily, together with empirical knowledge that water is actually H_2O" (Cohnitz & Häggqvist, 2018, p. 420). Again, we need knowledge for which properties it is the case that having them implies having them necessarily.

[30] Let me flag that *if* this works, then it is *not* the imagination that is doing the significant work, but whatever it is that provides us with the conflicting intuition. Given the number of cases we can generate, this might seem problematic in and of itself. As Kung notes, "[i]f everything ultimately hinges on a modal intuition, then the imagined situation is irrelevant" (2010, p. 651). I will leave this objection aside. The same goes for the other additional conditions: if they were to work, it is *not* the imagination that is doing the significant work.

[31] Remember our discussion from Chapter 1 (Section 1.2.2).

[32] For example, "I am in a position to develop a positive account of when imagination *does* provide evidence for possibility" (Kung, 2010, p. 637, original emphasis).

necessity is clearly a methodological non-starter. Hale (2003) forcefully warned us against this when he made the distinction between symmetric- and asymmetric epistemologies of modality and this worry has since been echoed throughout the literature. For example, Roca-Royes (2017) points out that, "[t]he methodological recommendation that emerges by reflecting on the issue of epistemic priority is as follows: aim at elucidating the *de re* possibility knowledge that we have about concrete entities in such a way that success here is *not* parasitic upon success in explaining knowledge of their essential facts" (p. 223, emphasis added). That is, "[w]e would like an account of a reliable, *autonomous* procedure for obtaining knowledge of [...] metaphysical possibility" (Hill, 2006, p. 230, emphasis added).

Independently of whether you think that focusing on an epistemology of possibility is correct, those who *do* aim to provide an epistemology of possibility should *not* rely on prior knowledge of necessities, as it would undermine their entire project.

3.6.1 Objection: A non-uniform epistemology of modality

One might suggest that maybe there is not such a strict separation between the epistemology of possibility and necessity. This seems to make the objections raised in this chapter less problematic.[33] However, even if there is no strict separation, there are a number of reasons to still consider the reliance on prior modal knowledge to be problematic for the QALC theorist.

First of all, remember that QALC imagination theorists promise to provide us with an account of how many of our ordinary possibility beliefs are justified. However, they have not delivered on this promise if all they do is push back the epistemological question to the epistemology of necessity and leave this unexplained. The explanatory value of the resulting QALC imagination-based epistemology would be incomplete and unsatisfactory as a philosophical explanation of the epistemology of possibility. To paraphrase Roca-Royes, as long as "such capacity for [necessity] knowledge is left unsatisfactorily explained, [...] this compromises (the satisfactoriness of) the elucidations they provide of our ordinary possibility knowledge" (2017, p. 244). This is, in a sense, a paraphrasing of methodological worry spelled out above, now aimed at those who think that there is no strict separation of or priority to the epistemology of possibility or necessity.

More importantly, the suggestion that there is no strict separation between the epistemologies of possibility and necessity misses the point. The QALC imagination theorists operate on the assumption that we can in fact imagine impossibilities (this is what sets them apart from the error-theorists) and still aim to provide at "a very reasonable epistemology of *possibility* [that] flows from a theory of imagination" (Kung, 2010, p. 621, emphasis added). Note that these theorists do not claim that all modal knowledge comes from imagination. In particular, they do not claim that knowledge of certain necessities (e.g., mathematical or logical) comes

[33] Thanks to Dominic Gregory and Margot Strohminger for useful discussions on these points.

from imagination. In that sense, they might agree with this objector that there is in fact a non-uniform epistemology of modality. However, when we focus on a particular group of claims (e.g., those pertaining to mundane possibilities claims about concrete objects), QALC imagination theorists hold that their account provides the justification for beliefs in possibility claims about them.[34] Suggesting that it is not so problematic to rely on prior knowledge of necessity when focussing on providing an epistemology of possibility is thus on a par with suggesting that the QALC imagination theorists should broaden their approach, allowing methods of the epistemology of necessity. In a sense, this is what I have been trying to argue, that the QALC imagination theorists *cannot* provide a significant epistemology of possibility (i.e., one that is able to deal with Kripkean *a posteriori* necessities) without letting in some prior knowledge of necessity.

3.7 Conclusion

Are there ways to avoid the problems raised in this chapter? QALC imagination theorists cannot, on pain of being QALC imagination theorists as opposed to error-theorists, reject the claim that we can imagine impossibilities. The whole point of their theories is to provide an imagination-based epistemology of possibility that incorporates the imaginability of Kripkean *a posteriori* impossibilities. In line with the assumptions of this dissertation (discussed in Chapter 1) we can also not accept radical scepticism nor move away from the focus on an epistemology of possibility.[35]

This leaves us with, roughly, two options: (i) we move away from the *QALC* imagination approach to imagination-based epistemologies of possibility or (ii) we move away from an *imagination*-based approach altogether. We can do the former by looking at *recreative* accounts of imagination, I will evaluate two such accounts in the following two chapters. In Part II, I will look at the latter option by discussing *similarity*-based epistemologies of possibility (see Hawke, 2011; Hartl, 2016; Hawke, 2017; Leon, 2017; Roca-Royes, 2017; Dohrn, 2019).

Either way, if what I have argued in here is correct, it will not be a QALC imagination-based approach that is the right explanation of our knowledge of modality.

[34] It seems that accepting a non-uniform epistemology for modal claims about concrete objects, is much harder to defend than suggesting that, overall, there might be a non-uniform epistemology of modality.

[35] Interestingly, there have been philosophers who, in response to similar troubles, have explicitly suggested the option to switch from an epistemology of possibility to an epistemology of necessity. For example, Crispin Wright, in recent work, seems to suggest something like this in light of worries presented by Bob Hale. Wright (2002) previously defended an epistemology of possibility, but in light of Hale's objections, he noted that one should maybe focus on an epistemology of necessity (Wright, 2018). In general, there is a growing literature on the necessity-first approach for epistemologies of modality (e.g., Hale, 2013; Jago, 2018; Kment, 2018; Mallozzi, 2018a; and Tahko, 2018).

Chapter 4

Pretense Imagination

[I]magination is a form—perhaps the central form—of conditional reasoning

– Langland-Hassan, 2016

In this chapter, we will discuss our first theory of imagination *as recreation*. In particular, we will discuss imagination as the recreation of *belief revision*, which we will call *pretense imagination*. This kind of imagination, as the name suggests, is used when we engage with pretense and fiction, but is also used for risk assessment, planning, and other cognitive phenomena. It is often argued that this kind of imagination is what justifies our beliefs in conditionals, which in turn are suggested to play a role in the epistemology of modality. In this chapter, I evaluate these claims.

I will first elaborate on a theory of what pretense imagination is. Then, I will discuss a formal model of pretense imagination. This formal model allows us to see very precisely where the justification through pretense imagination is supposed to come from. I will argue that pretense imagination justifies beliefs in conditionals only under certain conditions. Despite this, I argue, pretense imagination cannot serve as a basis for the epistemology of possibility, as some have argued.

4.1 Pretense: A Short Introduction

Consider the following experiment, which features example of the phenomenon known as *pretense*:

> The child is encouraged to 'fill' two toy cups with 'juice' or 'tea' or whatever the child designated the pretend contents of the bottle to be. The experimenter then says, 'Watch this!', picks up one of the cups, turns it upside down, shakes it for a second, then replaces it alongside the other cup. The child is then asked to point at the 'full cup' and at the 'empty cup' (both cups are, of course, really empty throughout).
>
> (Leslie, 1994, p. 223)

Children from as young as two years old already consistently point to the cup *that has been turned upside down* when asked to point at the 'empty cup' (see Leslie, 1994; Nichols & Stich, 2003). This suggests that children, at a very young age, are able to engage in pretense even if it goes against what they believe the world to actually be like.[1] One of the main questions that arises is *how* we develop a pretend scenario that seems so rational, but is often in contradiction with our explicit beliefs: the children explicitly believe that both cups are empty, yet they behave in pretense in a rational way *as if* one of the cups is full. They imagine this non-actual scenario in a *reality-oriented* way. Which logical rules, if any, govern the development of such a pretense scenario?

In this chapter, I will provide a formal model of, what we will call, *pretense imagination* by using tools from dynamic epistemic logic and belief revision theory. In this section and the next, I will review some of the current theories of pretense imagination and point to the essential features of pretense several theories agree on.[2] This is to let the formal model be informed, empirically and conceptually, by the current theories of pretense from the philosophy of imagination. In Section 4.3, I introduce branching-time belief revision structures in which the target notion of pretense imagination and a related notion of belief are formalised.[3] In Sections 4.4-4.7, I will evaluate the claim of some that pretense imagination plays a crucial role in the epistemology of possibility, through its role in the epistemology of conditionals (e.g., Williamson, 2007; Langland-Hassan, 2016; Williamson, 2016a). I conclude by discussing some potential objections to the arguments of this chapter in Section 4.8.

[1] The experiments confirm that the children believe that the cups are actually both empty.

[2] I will use 'imaginative episode,' 'imagination,' 'pretense', 'pretense imagination', and verbs such as 'imagining' as referring to the same kind of cognitive process for the purposes of this chapter, and this chapter only.

[3] For the purposes of this chapter, I will present a simplified version of a full model of pretense imagination, which would also capture the *aboutness* of pretense imagination. The full model is presented in Appendix B, with the philosophical aspects discussed in Section B.1 and the full logical models in Section B.2.

4.1.1 Imagination and Belief

As discussed in Chapter 2 (Section 2.1.1), the notion of 'imagination' is highly ambiguous and used in many different ways Kind, 2013; Balcerak Jackson, 2016; Liao & Gendler, 2019. In this chapter, we study the kind of imagination that is involved in pretense and pretend play, e.g., the kind of imagination used in the tea-party example from the previous page (following authors such as Currie & Ravenscroft, 2002; Nichols & Stich, 2003; and Langland-Hassan, 2012, 2016). I call it *pretense imagination*. This kind of imagination is not only used in pretense, but is also crucial for future planning, risk assessment, and other cognitive phenomena (see Byrne, 2005; Gopnik & Walker, 2013; Kind, 2016a; Lane et al., 2016). In this subsection, I discuss the relation between pretense imagination and beliefs; in particular, I look at several important ways in which beliefs restrict pretense imagination.

Consider again the example of the pretend tea-party, as described above. One thing we noted is that participants act in a *reality-oriented* manner with respect to which cups are full in the pretense and which are not. Similarly, when asked, *in the pretense*, where there is a puddle of tea after a full cup is held upside-down, it would be odd if the subject answers 'the ceiling', whereas it seems very natural to answer that there is a puddle on the floor. These two examples illustrate that pretense imagination is *restricted* in important ways by *belief*. Pretense imagination seems to follow *belief-like* patterns, which explains the rational, reality-oriented behaviour with respect to which cups are full. It also seems to be the case that *background beliefs* are imported into the imaginative episode, which explains the beliefs about, e.g., the workings of gravity in the pretense.[4] I will discuss these in turn.

One of the most prominent theories of pretense – that of Nichols & Stich (2003) – suggests that pretense reasoning is a cognitive capacity *functionally* the same as belief: pretense is *belief-like* reasoning. In other words, they argue that reasoning in pretense involves the same rational inference system that is deployed in actual reasoning about our beliefs (Nichols, 2006a). Similarly, Langland-Hassan (2012) – whose theory inspired the formal framework of this chapter – argues that pretense is reasoning about/with actual beliefs, but a very particular kind of belief.[5] He also points out that "imagination is a form—perhaps the central form—of conditional

[4] It has to be noted here that there are ways in which one can imagine recalcitrant situations with respect to both of these restrictions, namely if the agent *explicitly intervenes*. I set this complication aside for now and address the details of this in the next section.

[5] In the literature concerning pretense and the relevant imagination involved, there is a debate between those claiming that there is nothing over and above the cognitive attitudes belief and desire that is needed to account for what is going on during pretend play (the use of 'desire' here is meant in a non-technical, pre-theoretical sense) and those claiming that there is a specific cognitive capacity, distinct from belief and desire, that is involved (a pretense- or imagination-attitude). The former is called the *Single Attitude* (SA) account and the latter is a *Distinct Cognitive Attitude* (DCA) account. For example, for Langland-Hassan (2012), who identifies himself as an SA supporter, imagination is just a special case of belief and desire, whereas for Nichols & Stich

reasoning" (2016, p. 81). This is why, in the pretense, our beliefs about which cups are full after being filled are the same as they *would be* if some actual cups would be filled. To capture this in the formalism, we focus on belief and belief revision, where the latter is of *hypothetical* nature hinting at real belief changes were the pretend scenario to be actual. In this sense, it is sufficient to use models and operators that describe a situation where the objective facts of the world do not change but only the belief state of the imagining agent changes. Such a belief revision process follows, roughly, Ramsey's (1929) famous pattern, as described here by Stalnaker:

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent).
>
> (1968, p. 102)

I resort to the rich literature in (dynamic) epistemic logic and belief revision theory for the required modelling tools (see Section 4.3).

Another important factor that restricts pretense imagination, apart from being belief-like in its development, are the agent's *background beliefs* about the actual world. As Williamson notes, "[o]ne's imagination should not be completely independent of one's knowledge of what the world is like" (2016a, p. 114). For example, in the above pretense scenario, the subjects continue the pretense with the imagining that tea falls downwards as opposed to upwards because they *import* their background beliefs about gravitational forces – that unsupported objects fall towards the centre of the Earth – into the pretense. I will use the phrase 'taking on board' to refer to those beliefs that the agent accepts (also) into the pretense and uses to further the pretend scenario. The agent takes on board, in the imagination, that when full cups are turned upside down their contents fall down.[6]

With the relation between pretense imagination and belief on the table, let us turn to *what* pretense imagination is, how it functions, and what its crucial features are. In the next section, I describe in detail what I take pretense imagination to

(2003), who are supporters of DCA, imagination is *belief-like*, in that it is functionally similar to, yet distinct from, belief and desire. Even though I mainly follow Langland-Hassan's presentation, accepting SA is not essential to the models presented here, as one could reformulate everything in terms of a DCA account. So, ultimately, the model and arguments of this chapter could be interpreted either way.

[6] Note that it seems obvious that the agent does not take all their background beliefs on board. Why is it that some other background beliefs, such as Paris being the capital of France, water being a transparent liquid, etc., are not taken on board? I argue that one of the reasons why the subject does not imagine Paris being the capital of France in the tea-party situation is simply that the capital of France is *off-topic* and *irrelevant* to the pretend tea-party. This suggests a natural way to separate the background beliefs that can be taken on board in the pretense from the ones that are not: we select the *relevant* background beliefs to import into pretense based on what they are about, in other words, based on their *topics* (see Berto, 2018a,b). This will be discussed in detail in Appendix B.

be. Though most of what is said there is taken from the work of Langland-Hassan (2012, 2016), the resulting general picture (and thus the model thereof) captures most theories of pretense (e.g., that of Currie & Ravenscroft, 2002 and Nichols & Stich, 2003) and is compatible with certain theories of imagination (e.g., that of Byrne, 2005 and Williamson, 2007).

4.2 Pretense Imagination

In pretense, for example in the tea-party case above, the entire episode is made up out of a number of (temporally) shorter instances: the pretending that the cup is being turned upside-down, the tea is being poured. These instances are all 'part' of the entire tea-party pretense. It seems obvious that some of these are explicitly 'intended' by the agent, while others, e.g., the tea falling to the floor after the cup being held upside-down, develop without any intentions from the agent. Also, it seems very reasonable to assume that the pretense is full of choices from the agent that might go beyond what usually happens at a tea-party; the agent might, for example, say: 'Oh, a butler comes in to join the party'.[7] Let's discuss these important features in turn.

Explicit Input

Take an *imaginative episode* – e.g., the pretend tea-party – to be a sequence of individual *imaginative stages* – e.g., pouring the tea; keeping the cup upside down, et cetera. An imaginative episode always starts with a particular input. Langland-Hassan (2016) argues that imaginative episodes start with an *intention* of the agent. The intention that starts the imaginative episode consists of two parts. On the one hand, the intention provides the proposition that starts the imaginative episode. This is the proposition that makes up the first stage in the sequence of imaginative stages. As Langland-Hassan points out, "our intentions may be relevant in *initiating an imagining*" (2016, p. 65, emphasis added). On the other hand, the intention seems to play a role in demarcating what the imaginative episode (as a whole) *is about*. He says, "[o]ne's top-down intentions are key to initiating an imagining—in, say, *determining its general subject-matter*" (2016, p. 67, emphasis added).

Let's use the term *input proposition* to refer to the former and *overall topic* to refer to the latter in order to keep these two components clearly separated (the latter will only be used in Appendix B). An input proposition and overall topic together form the *explicit input* of an imaginative episode.

[7] See also Nichols & Stich (2003, pp. 23-24) for empirical evidence that people do make such choices in pretense.

Internal Development

Given an explicit input, the imaginative episode unfolds. In the case of the pretend tea-party, the development of this kind of imagination seems to follow a pattern that is very similar to that of rational belief revision. As Langland-Hassan puts it: "imagination [. . .] has its own norms, logic, or algorithm that shapes the sequence of i_x after the initiation of an imagining by a top-down intention" (Langland-Hassan, 2016, p. 67). The development of the imaginative episode is governed by the very same mechanisms that guide the inferences we make in rational belief updates (see Byrne, 2005; Nichols, 2006a; Williamson, 2007; Langland-Hassan, 2016; Williamson, 2016a; Berto, 2018b).[8] Let's call this kind of development the *internal development* of the imaginative episode. In terms of the tea-party example, this development makes the agent *automatically* take on board that the tea falls towards the ground when the cup is turned upside down. This nicely allows us to explain some of the features of imagination relating to the *reality-oriented* development of pretense. Moreover, the involuntariness of this step explains the non-arbitrary nature of imagination: we are not free to imagine whatever we want given a certain input and topic, which is supposed to render such mental exercises cognitively valuable (Byrne, 2005; Kind, 2016a; Balcerak Jackson, 2018).

Cyclical Interventions

Imagination, some have argued, is likely to have evolved in order to test a variety of actions to determine which one would be best to perform without having to actually perform the action and undergo all the risks that come with it (Nichols, 2006a; Langland-Hassan, 2016; Pezzulo & Cisek, 2016; Williamson, 2016a). "[T]here is much to be said," Langland-Hassan points out, "for the idea that imagination allows us to audition a *variety* of ways things might go, in order to choose a best course of action" (2016, p. 72, original emphasis). However, how can this be if all we have is the internal development of imagination? If that is all that we have, then given an input p in a situation s, we would expect that the outcome is always the same, namely whatever the result of a rational belief update with p in s is. This way, we can never test the variety of options given p in s through imagination. This is what Langland-Hassan dubs *the problem of deviance.*[9]

One way to think about how these variations occur is that the agent actively *intervenes* into the imaginative episode.[10] They add additional content forcefully

[8] Note that this is also accepted by those who disagree with the Nichols and Stich-like accounts of imagination (see Van Leeuwen, 2011 and Langland-Hassan, 2016).

[9] Nichols & Stich (2003) also note this problem and add a mechanism called *the Script Elaborator* to their sketch of a cognitive architecture. This Script Elaborator is supposed to fill in the details of certain familiar, or stereotypical, situations with details that go over and above the inferences that can be drawn from the content of the situations. However, as Langlang-Hassan points out, this is not so much a solution, as a label for the issue; leaving as much unexplained as before.

[10] Let me stress that we do not explicitly capture the *agentive* aspect of actively intervening.

(in that it does not necessarily follow from the previous imaginative stage) and this content can go beyond what the agent otherwise would have imagined. So, when testing the variety of potential outcomes given p in s, the agent actively intervenes somewhere in the imaginative episode with additional contents (e.g., q_1, q_2, etc.). Or, in Langland-Hassan's phrasing, imagination allows us to test a variety of actions "because we have *intentionally intervened in that processing.* To intentionally intervene is to stop the [internal development] where it is and to insert a new initial premise [. . .] into the [imaginative episode] for more processing" (2016, p. 74, emphasis added).[11]

From this discussion, toward a more systematic approach, I distil the following central features of a theory of pretense that we intend to capture in our formal framework.[12]

ROI: Pretense involves a form of *reality-oriented imagination.* The imagination involved in pretense is the kind that is, in a sense, restricted by (known) causal laws and that is the same as the imagination that is used to evaluate certain conditionals (e.g., 'what would happen if. . . ') (Byrne, 2005; Williamson, 2007).

PI: The imagination involved in pretense is strictly *propositional imagination.* That is, imagining *that* such and so is the case (Langland-Hassan, 2016). This is opposed to, e.g., sensory imagination (Gregory, 2019) or objectual imagination (Balcerak Jackson, 2018).[13]

ESP: The pretense always has an *explicit starting point.* This can either be in the form of an explicit external input ('Let's imagine that. . . ') or activated by something that caught the imaginer's attention (e.g., looking at an air plane might start off an imaginative episode where one pretends to be able to fly) (Langland-Hassan, 2016).

For logics that focus more on this action part of imagination see Wansing (2017), Olkhovikov & Wansing (2018, 2019), and, to some extent, Canavotto et al. (2020).

[11] For those who worry about phenomenology of an imaginative episode and the lack of 'active choice' that seems to be involved, note that most of this intervening happens sub- or unconsciously.

> What we might pre-theoretically think of as a single imaginative episode could in fact involve many such top-down 'interventions.' These interventions would allow for the overall imagining to proceed in ways that stray from what would be generated if one never so intervened. (Langland-Hassan, 2016, pp. 74-75)

[12] These features are compatible with most work on pretense and imagination (e.g., Currie & Ravenscroft, 2002; Nichols & Stich, 2003; Langland-Hassan, 2012, 2016; Berto, 2018b).

[13] This is not to say that there is no imagery involved in pretense, what I mean is that the kind of imagination that allows us to explain the pretense behaviour is propositional imagination.

QU: A crucial feature is, what has been called, *quarantining*. Pretense does not interfere with one's actual beliefs. One can pretend that p is the case irrespective of whether they believe that p or not-p (Nichols & Stich, 2003; Langland-Hassan, 2012).

RAT: Within the pretense, the agent reasons/behaves rationally; pretense seems to follow a 'belief-like' inference pattern (Nichols & Stich, 2003; Williamson, 2007; Langland-Hassan, 2012, 2016; Williamson, 2016a).

CHO: "[P]retence is full of *choices* that are not dictated by the pretence premise, or by the scripts and background knowledge that the pretender brings to the pretence episode [. . .] these choices typically get made quite effortlessly" (Nichols & Stich, 2003, p. 35).[14]

So, pretense imagination is a particular kind of *recreativist imagination*, namely the recreation of rational belief revision. In the next section, we take a first step towards a full-blown formal model of the logical development of pretense imagination.

4.3 Logic of Pretense Imagination

I propose a formal model of pretense imagination from which we can read off sequences of individual imaginative stages, denoted by $(i_1, \ldots, i_n)$, that form an imaginative episode, $\mathcal{I}$. As the pretense imagination follows 'belief-like' inference patterns and develops in stages, we'll use a simplified version of *branching-time belief revision models* introduced by Bonanno (2007). These models "provide a way of modeling the evolution of an agent's beliefs over time in response to informational inputs" (Bonanno, 2012, p. 206). In our framework, the imagined propositions will play the role of 'informational inputs'.[15]

4.3.1 Syntax and (idealised) Semantics

Let $\mathsf{Prop} = \{p_1, \ldots, p_n\}$ be a finite set of propositional variables and $\mathcal{L}$ be the language of classical propositional logic defined on Prop. The language $\mathcal{L}_{\mathsf{BI}}$ of the logic of belief and imagination is then defined by the grammar:

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B\psi \mid I\psi$$

[14] 'Pretense' is usually used to denote the imaginative episode *in combination with* the appropriate physical actions. So, in the case of the tea-party, when one moves their arm in the motion as if sipping tea from an empty cup, this is part of (and often the defining part of) the pretense episode. However, for our purposes, we ignore this part and only focus on the propositional imagination that is involved in such pretense.

[15] In Appendix B.2, we enrich these structures with a topicality component, following Berto (2018a,b), in order to render the imagining agent in question non-omniscient with respect to what they believe and imagine.

where $p \in \mathsf{Prop}$ and $\psi \in \mathcal{L}$. We read '$B\psi$' as 'the agent believes that ψ' and '$I\psi$' as 'the agent imagines that ψ'. It is important that we allow B and I to range only over Booleans. That is, our language $\mathcal{L}_{\mathsf{BI}}$ of belief and imagination follows the cognitive science and philosophy literature on imagination in focusing on *first-order* attitudes (e.g., Currie & Ravenscroft, 2002; Nichols & Stich, 2003; Byrne, 2005; Williamson, 2007; Langland-Hassan, 2012, 2016).[16] Finally, we employ the usual abbreviations for propositional connectives $\vee, \Rightarrow, \Leftrightarrow$ and we define $\top$ as $p \vee \neg p$ for some $p \in \mathsf{Prop}$, and $\bot := \neg\top$.[17]

We interpret $\mathcal{L}_{\mathsf{BI}}$ in (a version of) branching-time belief revision models, the first component of which consists in a forward-looking branching-time structure.

Definition 1. **Next-time Branching Frame**
A *next-time branching frame* is a pair $\langle S, \rightarrowtail \rangle$, where S is a non-empty set of *stages* and $\rightarrowtail$ is a binary relation on S such that for all $s_1, s_2, s_3 \in S$,

1. if $s_1 \rightarrowtail s_3$ and $s_2 \rightarrowtail s_3$, then $s_1 = s_2$ (no branching to the past);
2. if $(s_1, \ldots, s_n)$ is a sequence in S such that $s_i \rightarrowtail s_{i+1}$ for every $i \in \{1, \ldots, n-1\}$, then $s_n \neq s_1$ ($\rightarrowtail$ is strictly a next time relation).

Bonanno (2007) calls the elements of S 'instants' or 'dates', however, I prefer to call them 'stages', as we think of them as possible imaginative stages in which the agent could be. We read $s \rightarrowtail s'$ as "s' is an *immediate successor* of s" or "s is *the immediate predecessor* of s'". Every stage has at most a unique immediate predecessor (see Definition 1.1), but can have several immediate successors. In particular, we'll use *rooted* next-time branching frames in order to explicitly mark the actual belief state of the agent. To define a rooted frame, we let $\rightarrowtail^+$ denote the transitive closure of $\rightarrowtail$. A next-time branching frame $\langle S, \rightarrowtail \rangle$ is *rooted* if there is $s_0 \in S$ such that $s_0 \rightarrowtail^+ s'$ for all $s' \in S$ with $s_0 \neq s'$. We call such an s_0 *the initial stage.*

Definition 2. **Branching-time Belief Revision Model**
A *branching-time belief revision model* (in short, *model*) is a tuple $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V \rangle$, where

1. $\langle S, \rightarrowtail \rangle$ is a *rooted next-time branching frame*;
2. W is a finite set of *possible worlds* or *states*;

[16] If one looks in, e.g., *The Routledge Handbook of Philosophy of Imagination* (Kind, 2016c), there is no mention of second- or higher-order imaginings.

[17] In this chapter, I use '$\Rightarrow$' for the *material* conditional. I do so because later we will be concerned with the *indicative conditional*, for which I reserve the '$\rightarrow$' symbol.

3. $\preceq: S \to W \times W$ is a function that assigns every stage in $s \in S$ a total preorder on W, denoted by $\preceq_s$;[18]

4. $V : \mathsf{Prop} \to \wp(W)$ is a valuation function that maps every propositional variable in Prop to a set of possible worlds.

S is the set of stages, which model the stages of the imaginative episode, and W is the set of worlds that model the agents beliefs at these stages. Because the purpose of this model is to make it precise how the reasoning of ordinary humans develops in pretense, I opt for simplicity and use models with finite worlds in this chapter.[19] '$\preceq_s$' is the *plausibility order at stage* s and represents the arrangement of worlds to the degree that the agent considers them plausible at s. We read '$w \preceq_s v$' as 'w is at least as plausible as v at stage s'. We say 'w and v are *equally plausible* at stage s', denoted by $w \approx_s v$, if $w \preceq_s v$ and $v \preceq_s w$. We define *strict plausibility*, denoted by $\prec_s$, in a usual way as $w \prec_s v$ iff $w \preceq_s v$ and $w \not\approx_s v$.

The set of minimal elements, $Min_{\preceq_s}(U)$, for any $U \subseteq W$ with respect to $\preceq_s$, is defined as

$$Min_{\preceq_s}(U) = \{w \in U : w \preceq_s v \text{ for all } v \in U\}.$$

The set $Min_{\preceq_s}(W)$ forms the set of possible worlds the agent considers *most plausible* at s. Since W is finite, every non-empty subset of W has a minimal element with respect to each $\preceq_s$, i.e., $Min_{\preceq_s}(U) \neq \emptyset$ for all $U \subseteq W$ such that $U \neq \emptyset$. As readers familiar with Dynamic Epistemic Logic might already have observed, for each $s \in S$, $(W, \preceq_s, V)$ constitutes a standard plausibility model (see Baltag & Smets, 2006; van Benthem, 2007).

For an illustration of a branching-time belief revision model, see Figure 4.1; the nodes in the figure represent the stages of the imaginative episode with the plausibility ordering of that particular stage – i.e., $\preceq_i$ represents the plausibility ordering at stage s_i.

Strictly speaking, these models capture *what an agent has imagined throughout an imaginative episode* as in 'what the agent has taken on board in the development until now' (where 'taken on board' can spelled out in two different ways, which will be explained further in Section 4.3.2). What an agent imagines in an episode is the cumulative content of the hypothetical belief revisions on a particular branch. The root of the model represents the stage the agent has not yet started the imaginative process and the branches of the model represent the possible ways the agent's imagination *can* develop. We can therefore see the plausibility structure *at the initial stage* as the model that represents the agent's *actual* doxastic state and their *pretend* or *simulated* doxastic states are represented by the further stages in a branching-time

[18] A total preorder $\preceq_s$ on W is a reflexive and transitive binary relation such that either $w \preceq_s v$ or $v \preceq_s w$ for all $w, v \in W$.

[19] In future work, we aim to provide a full-fledged logical account including models with possibly infinite worlds, but this is beyond the scope of this chapter.

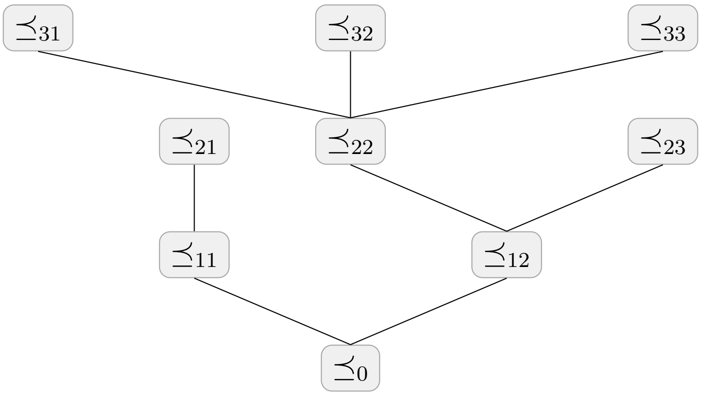


Figure 4.1: *An example of a branching-time belief revision model.*

belief revision model.[20] Remember the tea-party example, the initial stage represents the actual beliefs of the children before they engage in the pretense; at this stage they believe both cups are empty. Then the pretense starts and the children *hypothetically* revise their beliefs as they would if the tea-party were actual; at some stage the children *hypothetically* believe that both cups are full.

This kind of imagination can be read off of the actual development of the hypothetical belief revisions in the pretense scenario, represented by a finite sequence of linear stages, called *history*, h:

$$h = (s_0, s_1, \ldots, s_n) \text{ such that } s_i \rightarrowtail s_{i+1},$$

where s_0 is the root of the underlying next-time branching frame. We call s_0 the *initial stage* and s_n the *current stage*. History h thus keeps track of the past stages, but does not tell us anything about the future. Given a branching-time belief revision model $\mathcal{M}$ and a history $h = (s_0, s_1, \ldots, s_n, s_{n+1})$, we will be able to extract the corresponding imaginative episode $\mathcal{I} = (i_1, \ldots, i_n)$ as described in Section 4.3.2.

Our agent gets from one stage to the next of a history by following a certain belief revision process. Here we choose to model agents who revise their beliefs according to the well-known *lexicographic upgrade* policy. This choice does not bear on substantive philosophical points and, in principle, one can employ other belief revision policies – such as the so-called conservative upgrade (see van Benthem, 2007) – in a similar way.

Definition 3. **Lexicographic Upgrade**
Given a pre-ordered set $\langle W, \preceq_s \rangle$ and $U \subseteq W$, the upgraded pre-order by U is the

[20] Let me emphasise that the only component of the model that varies from stage to stage is the plausibility ordering and that the valuation of the propositional variables stays the same throughout the stages of a branching-time structure. This means that our branching-time structures represent simulated belief changes of the imagining agent in a world where the objective facts do not change.

tuple $\langle W, \preceq_s^U \rangle$, where $\preceq_s^U$ is the new ordering such that $w \preceq_s^U v$ iff (1) $w \preceq_s v$ and $w \in U$, or (2) $w \preceq_s v$ and $v \in W \backslash U$, or (3) ($w \preceq_s v$ or $v \preceq_s w$) and $w \in U$ and $v \in W \backslash U$.

In other words, upon receiving information U, lexicographic upgrade makes all U-worlds strictly more plausible than all $W \backslash U$-worlds and keeps the ordering the same within those two zones (van Benthem, 2007, p. 141).

We now have the required tools to give the semantics for our language. Formulas of $\mathcal{L}_{\mathsf{BI}}$ are evaluated not only with respect to states, but with respect to state-history pairs of the form $\langle w, h \rangle$. Thus, the *intension of* φ with respect to h in $\mathcal{M}$ is $|\varphi|^h_{\mathcal{M}} := \{w \in W : \mathcal{M}, \langle w, h \rangle \Vdash \varphi\}$ (I omit the subscript $\mathcal{M}$ and superscript h when the model and actual history are clear from the context). I will adopt the following notation for convenience: $\preceq^{\varphi}_{s_k} = \preceq^{|\varphi|^h_{\mathcal{M}}}_{s_k}$ and $h[k] = (s_0, \ldots s_k)$ is the initial segment of h of length $k+1$.

Definition 4. **$\Vdash$-Semantics for $\mathcal{L}_{\mathsf{BI}}$**
Given a model $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V \rangle$ and a world-history pair $\langle w, h \rangle$ such that $h = (s_0, s_1, \ldots, s_n)$, the semantics for $\mathcal{L}_{\mathsf{BI}}$ is defined recursively as follows:

$$
\begin{array}{lll}
\mathcal{M}, \langle w, h \rangle \Vdash p & \text{iff} & w \in V(p) \\
\mathcal{M}, \langle w, h \rangle \Vdash \neg\varphi & \text{iff} & \text{not } \mathcal{M}, \langle w, h \rangle \Vdash \varphi \\
\mathcal{M}, \langle w, h \rangle \Vdash \varphi \wedge \psi & \text{iff} & \mathcal{M}, \langle w, h \rangle \Vdash \varphi \text{ and } \mathcal{M}, \langle w, h \rangle \Vdash \psi \\
\mathcal{M}, \langle w, h \rangle \Vdash B\varphi & \text{iff} & Min_{\preceq_{s_n}}(W) \subseteq |\varphi|^h_{\mathcal{M}} \\
\mathcal{M}, \langle w, h \rangle \Vdash I\varphi & \text{iff} & \exists k < n(\preceq_{s_{k+1}} = \preceq^{\varphi}_{s_k} \text{ and } \mathcal{M}, \langle w, h[k+1] \rangle \Vdash B\varphi)
\end{array}
$$

For any $\Sigma \subseteq \mathcal{L}_{\mathsf{BI}}$ and $\varphi \in \mathcal{L}_{\mathsf{BI}}$, φ is said to be a *logical consequence* of Σ, denoted by $\Sigma \vDash \varphi$, if for all models $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V \rangle$ and all world-history pairs $\langle w, h \rangle$ of $\mathcal{M}$: if $\mathcal{M}, \langle w, h \rangle \Vdash \psi$ for all $\psi \in \Sigma$, then $\mathcal{M}, \langle w, h \rangle \Vdash \varphi$. For single-premise entailment, we write $\psi \vDash \varphi$ for $\{\psi\} \vDash \varphi$. As a special case, *logical validity*, $\vDash \varphi$, truth at all world-history pairs of all models, is $\emptyset \models \varphi$, entailment by the empty set of premises.

It is not difficult to see that the truth of Booleans in a given model is stage and history *independent*, that is, their truth values depend only on the actual world.

Lemma 1. For every model $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V \rangle$, world-history pairs $\langle w, h_1 \rangle$ and $\langle w, h_2 \rangle$ in $\mathcal{M}$, and $\varphi \in \mathcal{L}$, we have $|\varphi|^{h_1}_{\mathcal{M}} = |\varphi|^{h_2}_{\mathcal{M}}$.

Proof. The proof is straightforward by subformula induction on φ. □

The intension of a Boolean φ in $\mathcal{M}$ can therefore be written as $|\varphi|_{\mathcal{M}} = \{w \in W : \mathcal{M}, \langle w, s_0 \rangle \Vdash \varphi\}$. Moreover, the truth of sentences involving only the *belief* modality do not depend on the whole history, but only on the actual world and current stage.

The belief modality therefore represents the agent's beliefs at a particular stage, that is, it cannot refer to the past stages in the given history. It is important to remember at this point that the belief modality represents the *hypothetical* or *pretense* beliefs of the agent throughout a history, except when $h = (s_0)$. The agent's *actual* beliefs are given by the plausibility model in the initial stage, s_0. We could add a specific modality that reflects the agents actual doxastic state as follows:

$$\mathcal{M}, \langle w, h \rangle \Vdash B_{@}\varphi \text{ iff } Min_{\preceq_{s_0}}(W) \subseteq |\varphi|^h_{\mathcal{M}}.$$

Imagination, on the other hand, is dependent on both w and the *whole history h*. According to the proposed semantics, an agent imagines φ if they have successfully revised their belief state with φ at some earlier stage in the history.[21] In other words, we take what an agent imagines at the current stage to be the cumulative collection of propositions by which they have upgraded their (simulated) belief state at some stage before the current one. A less terse and more appropriate reading of $I\varphi$, then, is that "the agent has taken φ on board at some stage of the imaginative episode". In this sense, the imagination operator I is a backward looking modality that keeps track of the informational input the agent uses through an imaginative episode. Moreover, although the agent never imagines $\bot$ (see footnote 21), due to the definition of lexicographic upgrade, nothing stands in the way of imagining φ while believing (either really or in the pretense) $\neg\varphi$, or imagining φ and imagining $\neg\varphi$ in one imaginative episode, as φ and $\neg\varphi$ can be taken on board at different stages.

4.3.2 Full Models and Imaginative Episodes

Note that the models described above are too liberal for the following reasons. We want to model agents who can in principle imagine whatever they believe at any stage and who revise their beliefs *only* according to the lexicographic upgrade policy described in Definition 3. While the semantic clause for $I\varphi$ is concerned only with the stages that the agent can reach via lexicographic upgrade, the model does not yet have any restrictions on the plausibility orderings at successive stages. Thus,

[21] The agent is said to have *successfully* revised their beliefs by φ at some stage s in the given history if they believe φ in the next stage. This corresponds to the *Success Postulate* of the AGM belief revision theory (Alchourrón et al., 1985) and, as B ranges only over Booleans, our framework is not subject to problems concerning higher-order beliefs such as the Moorean phenomena (Holliday & Icard, 2010). Due to the second conjunct in the semantic clause of $I\varphi$ in Definition 4 (that is, $\mathcal{M}, \langle w, h[k+1] \rangle \Vdash B\varphi$), our imagination operator is always concerned with the so-called successful revisions (for the sake of brevity, I usually drop the phrase "successful"). In fact, lexicographic upgrade by definition always leads to successful revisions as long as the intension of the new informational input is non-empty. Since $Min_{\preceq_s}(W) \neq \emptyset$ for all s in every model, $\neg B\bot$ is a validity with respect to the proposed semantics. This means that the agent never believes (actually or in pretense) nor imagines blatant contradictions (where the latter is guaranteed by the above mentioned component in the semantic clause of $I\varphi$).

the agent can *potentially* follow any belief revision policy (in the sense of any way of changing the plausibility ordering). We therefore need to impose further restrictions to obtain our intended models. Let's call such restricted models *full models*, which are defined below.

Definition 5. A *full model* $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V\rangle$ is a branching-time belief revision model such that,

1. for all $w \in W$, $h = (s_0, s_1, \ldots, s_n)$, and $\varphi \in \mathcal{L}$, if $\langle w, h\rangle \Vdash B\varphi$, then there is an $s' \in S$ such that $s_n \rightarrowtail s'$ and $\preceq_{s'} = \preceq^{\varphi}_{s_n}$;
2. for all $s, s' \in S$, if $s \rightarrowtail s'$ then $\preceq_{s'} = \preceq^{\varphi}_{s}$ for some $\varphi \in \mathcal{L}$.

The first condition says that if the agent believes φ at w with respect to history h, then there is a possible next pretend belief stage the agent could reach by lexicographically upgrading their beliefs with φ. This condition guarantees that the agent can take on board/imagine whatever they believe at any stage.[22] The second condition states that the agent revises their beliefs only according to the lexicographic upgrade policy: if s' succeeds s, the plausibility order $\preceq_{s'}$ is a lexicographic upgrade of the plausibility order $\preceq_s$.

Internally Developed and Intervened Imaginative Content

Recall that Langland-Hassan (2016) distinguishes between imaginative stages that follow internally from their predecessors and those that are added through intervention (see Section 4.2). Our model allows us to capture this distinction very nicely. Given a history $h = (s_0, \ldots, s_n)$ and $k \leq n$, recall that $h[k] = (s_0, \ldots s_k)$ is the initial segment of h of length $k+1$. We then define the kth imaginative stage i_k, *the set of sentences the agent has imagined up to stage* k, as

$$i_k = \{\varphi \in \mathcal{L} : \langle w, h[k]\rangle \Vdash I\varphi\}$$

This way we extract the imaginative stages through the actual history and define the corresponding imaginative episode $\mathcal{I} = (i_1, \ldots, i_n)$ as a sequence of sets of sentences in $\mathcal{L}$. It is not difficult to see that $i_0 = \emptyset$ as the semantics of I refers to strictly earlier stages than the current one (see Definition 4). An imaginative episode starts with an input proposition, forming the first imaginative stage i_1 and then develops into the full imaginative episode. In order to distinguish between stages that follow through internal development and stages that are added through intervention, let's introduce two distinct operators into our language: $I_i\varphi$ and $I_a\varphi$. The former

[22] Moreover, since $B\top$ is a validity, we have that for all $s \in S$ there is an $s' \in S$ such that $s \rightarrowtail s'$ and $\preceq_s = \preceq_{s'}$. This in particular means every stage has at least one successor which is a copy of itself.

concerns *i*nternally developed stages and the latter concerns *a*dded content through intervention. These two modalities are interpreted in full models as follows:

$$\begin{array}{lll} \langle w,h\rangle \Vdash I_i\varphi & \text{iff} & \exists k<n((\preceq_{s_{k+1}}=\preceq^{\varphi}_{s_k} \text{ and } \langle w,h[k+1]\rangle \Vdash B\varphi) \text{ and } \langle w,h[k]\rangle \Vdash B\varphi)^{23} \\ \langle w,h\rangle \Vdash I_a\varphi & \text{iff} & \exists k<n((\preceq_{s_{k+1}}=\preceq^{\varphi}_{s_k} \text{ and } \langle w,h[k+1]\rangle \Vdash B\varphi) \text{ and } \langle w,h[k]\rangle \nVdash B\varphi) \end{array}$$

Semantically, $I_i\varphi$ states that 'the agent takes φ on board at some stage of the actual history *where they already believe it*'. The proposition expressed by φ is in this sense part of the internal development. The agent does *not* have to add to the imaginative episode *everything she believes* at a certain stage but they further the imaginative episode via *some* of the already believed propositions. On the other hand, $I_a\varphi$ says that 'the agent takes φ on board at some stage of the actual history and φ was not believed at that stage'. This implies that φ was imagined not as a result of the agent's belief revision process, but added 'externally' to the imaginative episode. The proposition expressed by φ is in this sense *added* content through intervention.[24] For example, when the cups are held upside down in the tea-party pretense, the imagination develops internally with, something like, 'the tea falls down'. Whereas when the child unexpectedly imagines that a giraffe comes to the tea-party (which I take not to 'follow from' beliefs about a tea-party), we say that the content is actively intervened.

Let me conclude this section by relating the model developed here back to the feature of pretense imagination as discussed in Section 4.2. We have focused particularly on propositional (hypothetical) belief and imagination, so **PI** requires no comments. **RAT** and **ROI** are accounted for partly because the development of an imaginative episode follows the belief revision policy *lexicographic upgrade*. Moreover, as we will see in Appendix B, imagination is restricted in important ways by the overall topic of the imaginative episode and the totality of topics the agent has grasped, making the formalised notion of imagination reality-oriented. For **ESP**, recall that we define the kth imaginative stage i_k as $i_k = \{\varphi \in \mathcal{L} : \langle w, h[k]\rangle \Vdash I\varphi\}$. The corresponding imaginative episode $\mathcal{I} = (i_1, \ldots, i_n)$ is then obtained from the actual history, where i_1 constitutes the explicit starting point of the imaginative episode. Moreover, the plausibility structure at the initial stage s_0 represents the agent's actual doxastic state and their pretend doxastic states are represented by the further stages in a branching-time belief revision model. So, throughout an imaginative episode, the actual beliefs of the agent are kept fixed and only the pretend beliefs are revised. This gives us **QU**. Finally, **CHO** is accounted for as our models are rich enough to distinguish the operators I_i and I_a, where the latter is concerned with added content through intervention.

[23] It is easy to see that the component '$\langle w, h[k+1]\rangle \Vdash B\varphi$' in the beginning of the semantic clause of $I_i\varphi$ is redundant: $\langle w, h[k]\rangle \Vdash B\varphi$ guarantees that $|\varphi|_{\mathcal{M}} \neq \emptyset$, thus, $\preceq_{s_{k+1}}=\preceq^{\varphi}_{s_k}$ implies that $\langle w, h[k+1]\rangle \Vdash B\varphi$, that is, lexicographic upgrade leads to successful revision by φ.

[24] Note, that the definition of $I_a\varphi$ only works as a sufficient condition for the content being added and leaves the possibility of intervention of already believed propositions open.

4.4 Epistemology of Pretense Imagination

With this formal model of pretense imagination at hand, we now turn to the *epistemology* of pretense imagination. Langland-Hassan (2016) and Williamson (2007, 2016a) both argue that pretense imagination might be central to conditional reasoning and the epistemology of conditionals. However, what we are interested in is the particular use of these conditionals to gain knowledge of possibilities and whether pretense imagination plays a crucial role in the epistemology of these conditionals (Williamson, 2007, 2016a). So, we need to evaluate two claims: (i) can pretense imagination provide justification for believing conditionals and (ii) can the pretense imagination, in virtue of (i), play a role in the epistemology of possibility?

The model presented in this chapter allows us to very precisely evaluate these claims. We saw that imaginative episodes are sequences of imaginative stages; these stages are *either* explicitly intervened by the agent *or* developed through hypothetical belief revisions. So, we can evaluate the epistemic usefulness of pretense imagination through an argument by cases. First, I will argue that internally developed imagination cannot be used to gain justification for new beliefs in conditionals, after which I will argue that particular instances of intervened content do give rise to new beliefs in conditionals. Despite this, I will conclude by arguing that still, pretense imagination cannot explain our knowledge of *non-actual possibilities*.

4.4.1 Beliefs in Conditionals and Conditional Beliefs

The first thing to stress is that I will focus on our beliefs in *indicative conditionals* (represented with '$\rightarrow$').[25] However, the logic discussed above does not involve an indicative conditional. In this subsection, I will first argue that we have in fact all we need to evaluate the *epistemological* question whether pretense imagination can provide us with justification for new beliefs in indicative conditionals.

A venerable tradition of how to determine whether we should believe a conditional has it that we should believe a conditional if we believe the consequent after having (hypothetically) revised our beliefs with the antecedent. This traces back to, at least, Ramsey, who suggested that if we are to determine 'If φ, then ψ' and we are uncertain about the antecedent, then we should add φ "hypothetically to [our] stock of knowledge" and then evaluate "on that basis" whether ψ (1929, p. 247, fn. 1). Stalnaker (1968) and Williamson (2007, ch. 5) also suggest epistemologies of conditionals in this vein.[26] If such theories are correct, then if the agent has a

[25] In the conclusion of this chapter, whether Langland-Hassan and Williamson focus on the indicative conditional and whether some of the arguments of this chapter carry over to *other* conditionals.

[26] For example, "one supposes the antecedent and develops the supposition. [...] To a first approximation: one asserts the counterfactual conditional if and only if the development eventually leads one to add the consequent" (Williamson, 2007, pp. 152-153). See also the quote from

rational conditional belief, then they are equally in a position to justifiably believe the corresponding conditional.

Even though we do not have an indicative conditional in our semantics, we do have everything we need to define *conditional beliefs* in our model. Given Definitions 2 and 3, we can define conditional beliefs in our models as follows:

Definition 6.

$$\mathcal{M}, \langle w, h \rangle \Vdash B^{\varphi}\psi \quad \text{iff} \quad Min_{\preceq^{\varphi}_{s_n}}(W) \subseteq |\psi|^{h}_{\mathcal{M}}, \text{ where } h = (s_0, \ldots, s_n)$$

We take '$B^{\varphi}\psi$' to be a *conditional belief*: the agent believes ψ given (or conditional on) φ. Epistemologically speaking, if something like the Ramsey-test is a correct epistemology of conditionals, a conditional belief is similar enough to a belief in the corresponding conditional for our purposes.[27]

Some Empirical Support

There is some empirical evidence that the epistemological relation between conditional beliefs and beliefs in conditionals, on which a Ramsey-test epistemology for conditionals relies, is true. That is, there is empirical evidence that suggests that people believe conditionals if they also have the corresponding conditional belief. In order to properly spell out the evidence and how it supports (something like) a Ramsey-test epistemology for conditionals, we need to say a bit more about the relationship between beliefs, acceptability, and probability.

I focus on empirical data from Douven and colleagues (Douven & Verbrugge, 2010; Douven, 2013; Douven & Verbrugge, 2013; Douven, 2015), but as Douven points out, the relevant empirical findings have been tested in many different forms, by many different researchers over the last decade (see Douven, 2013, p. 11, fn. 10; Douven & Verbrugge, 2013, p. 712; and Elqayam & Over, 2013 for references to this empirical literature). The empirical data focuses on, what in the psychology of reasoning literature is known as, *the Equation* (EQ): the subjective probability (or the degree of belief) of a conditional 'if φ then ψ' is the corresponding (subjective) conditional probability $\texttt{Pr}(\psi|\varphi)$ (where '$\psi|\varphi$' means 'ψ given φ').[28] That is, where '$\texttt{Pr}(\varphi)$' is the subjective probability or degree of belief in φ, $\texttt{Pr}(\psi|\varphi) = \texttt{Pr}(\varphi \to \psi)$. This suggests "that people evaluate the probability of conditionals as the conditional probability for a wide range of conditionals" (Elqayam & Over, 2013, p. 253).[29] Importantly, as already mentioned, this epistemological claim has,

Stalnaker (1968, p. 102) on page 62.

[27] I say 'similar enough' here because, as we will see below, the conditional belief and belief in the conditional occur at *different* stages in the model.

[28] Note that this might only hold for 'simple conditionals' – i.e., conditionals $\varphi \to \psi$ where 'φ' and 'ψ' do not contain any conditionals themselves. As we also focus on first-order imagination, this limitation is not a problem for us.

[29] Note that we are *not* interested in the question of whether these things are the same *mental*

> over the past decade, [...] been subjected to empirical testing by various experimental psychologists, and it has been found, time and again, that people's judged probabilities of conditionals do closely match their judgments of the corresponding conditional probabilities [...] Given these experimental results, rejecting (EQ) would amount to attributing massive error to people as far as their judgments of [...] conditionals are concerned. (Douven & Verbrugge, 2013, p. 712)

All these empirical tests show that there is an epistemological equivalence in terms of *subjective probabilities* in conditionals and conditional subjective probabilities; in that "people do generally judge the probability of a conditional to be equal to the corresponding conditional probability" (Douven, 2013, p. 11).

These data concern the subjective probabilities of agents (i.e., it is *quantitative*), whereas our definition of conditional belief is defined as belief *tout court* (i.e., it is *qualitative*). So, in order for the data to support the use of Definition 6 in the epistemology of conditionals, we need to find a way to make the data on the quantitative epistemological equivalence relevant to the qualitative relationship between conditional beliefs and beliefs in conditionals. We start with the *Lockean Thesis* (LT) (Foley, 1992):

(LT) "A proposition φ is acceptable iff $\texttt{Pr}(\varphi) > \theta$", where '$\theta$' is some threshold. (Douven, 2015, p. 103)

Secondly, we need to relate acceptability to belief. Douven provides us with a straightforward way of doing so.[30]

> As I understand the term 'acceptability,' it designates justified or rational believability. To say that a given proposition is acceptable for a person is to say that it is epistemically all right for the person to adopt that proposition as a belief. (Douven, 2015, p. 91)

Combining this with (LT), we get what I will call the *Lockean Thesis for Belief* (LTB):[31]

states (see, e.g., Leitgeb, 2007). Additionally, one might worry that this gives rise to the famous *triviality results* (Lewis, 1976; Gärdenfors, 1988). However, given that we allow our belief- and imagination-operators to range only over Booleans, these triviality worries do not seem to apply. See Douven (2013) for a discussion about the tension between the empirical findings and the formal triviality results.

[30] Not everyone would define acceptability in the way that Douven does, see for example Engel (1998). However, given that we are working with Douven's data, I take it to be unproblematic to use his interpretation.

[31] Many have indeed focused on this belief-version of the Lockean Thesis (e.g., Foley, 1992; Hawthorne, 2009; Demey, 2013).

(LTB) A proposition φ is rationally believable for a person iff $\mathtt{Pr}(\varphi) > \theta$, where '$\theta$' is some threshold.

In what follows, I will assume that the threshold is fixed and suppress any mention of it. We can now link subjective probabilities that agents assign to propositions to qualitative beliefs by using (LTB). Going back to the data of Douven and colleagues, we can replace the subjective probabilities with the corresponding beliefs in (EQ) – as per (LTB). For the purposes of this chapter, this means that, instead of talking about 'judge' or 'evaluate', we can say the following:

> If people conditionally believe ψ given φ, they are also in the epistemological position to justifiably believe the conditional 'if φ then ψ'.

This suggests that Definition 6 is enough to evaluate the claim that pretense imagination provides us with justification for new beliefs in conditionals.

Let me stress that these empirical findings are supposed to *support* the philosophical claim that the epistemology of (indicative) conditionals relies on hypothetically revising one's beliefs with the antecedent of the conditional and checking to see if one ends up (hypothetically) believing the consequent. Additionally, I should point out that the empirical data often is of the form 'there is *no significant difference* between judgements of conditional beliefs/probability and the belief in/probability of the conditional'. However, generally one should not conclude that there is no difference between two judgements based on results showing that no significant difference is found. The reason why, in this case, we can still take the empirical findings to be in support of the relation between conditional beliefs and beliefs in conditionals is because of the number of the empirical findings. As Douven (2013, pp. 12) points out, throughout the literature, the Equation has been tested in many different ways, shapes, and forms and almost always the results were the same. This is abductive evidence that, despite the fact that we cannot statistically secure an equivalence, the lack of a significant difference does suggest that there is no difference.[32]

Now that it is plausible that we can use Definition 6 to evaluate the claim that pretense imagination plays a role in the epistemology of conditionals and of possibility, let's turn to discuss the epistemic usefulness of the internal development and the intervened content in turn.

4.5 Epistemic Usefulness of Internal Development

Both Langland-Hassan (2016) and Williamson (2016a) suggest that what makes imagination epistemically useful is that it is able to go beyond the agent's intentions.

[32] To further strengthen these empirical results, we should look at the *statistical power* of these results in combination with their sample size. In order to overcome the limitations of individual studies, ideally a meta-analysis would be performed on the empirical data concerning the Equation. As far as I am aware, no such a meta-analysis has yet been done.

For example, after "having forced the initial conditions, [the imaginer] lets the rest of the imaginative exercise unfold without further interference. For that remainder, his imagination operates in involuntary mode" (Williamson, 2016a, p. 116) and "surprise may come in the influence of the lateral algorithms [i.e., what is called the 'internal development' above] themselves. They are what take the imagining beyond one's intentions" (Langland-Hassan, 2016, p. 76). In our model, the kind of 'involuntary' development which takes the imagination beyond the agent's intentions is captured by the internal development, so it seems plausible that this is where pretense imagination gets epistemological force. In particular, the idea that both Williamson and Langland-Hassan seem to defend is that after an imaginative episode with explicit input φ, if you end up at some point imagining ψ, the knowledge that you gain is of the (indicative) conditional $\varphi \to \psi$ (see also Nichols & Stich, 2003; Byrne, 2005; and a lot of the suppositional reasoning literature following the Ramsey test). "[T]he inferences drawn in imagination are [then] imported back into one's beliefs as consequents to a *newly believed* conditional" (Langland-Hassan, 2016, p. 68, emphasis added).

I will argue that the conditionals that one, after internal development, might import back into one's beliefs are not *new*ly believed conditionals.

Given the definition of conditional beliefs in our model (Definition 6) and the argument that this is epistemologically enough for beliefs in conditionals, we can properly evaluate the claim that the internal development of pretense imagination justifies new beliefs in conditionals. Before we give the precise formulation of the claim we set out to evaluate, it is important to stress that in this part of the argument by cases, we focus only on *internal development*: an imaginative episode where we *only* rely on hypothetical belief revision with (hypothetically) believed propositions. That is, for any world-history pair that we consider here, the history, $h = (s_0, \ldots, s_n)$, is such that for any $i < n$, $\preceq_{s_{i+1}} = \preceq^{\varphi}_{s_i}$ for some $\varphi \in \mathcal{L}$ such that $\langle w, h[i] \rangle \Vdash B\varphi$. Let us call such a history an *internally developed history.*

The claim that the internal development of pretense imagination can provide us with justification for *new* conditional beliefs is as follows: it is possible to come to have a conditional beliefs somewhere in an internally developed history such that the agent *does not* have that conditional belief at the root stage – i.e., the conditional belief is new. For if revising one's beliefs with the antecedent results in believing the consequent *and* the conditional was not yet believed in the original state of the imaginer, then it can be said that the imaginative episode provided the justification for a *new* belief in the conditional. Call this the *target claim.*

I will argue that this is *false* – i.e., I will argue that the beliefs that are the result of such imaginative episodes are not *new* beliefs.

To show that the target claim is false, I will prove that for any internally developed history, if there is a stage where revising one's beliefs at that stage with φ results in believing ψ at the next stage, then the agent *already* had a conditional belief in ψ

given φ in the root stage – i.e., the conditional belief is not new.

Show: For all internally developed histories, $h = (s_0, \ldots . s_n)$, and all formulas φ and ψ, *if* there is an $i < n$ such that $\preceq_{s_{i+1}} = \preceq^{\varphi}_{s_i}$ and $\langle w, h[i+1] \rangle \Vdash B\psi$ *then* $\langle w, s_0 \rangle \Vdash B^{\varphi}\psi$.

The first thing to note is a consequence of our belief revision policy and the effects this has for an internally developed history. Remember that when we revise our beliefs with φ, the plausibility ordering *amongst* all the φ-worlds remains the same (see condition (1) of Definition 3).[33] Thus, upgrading our beliefs with a believed proposition – i.e., a proposition such that it is true at all the most plausible worlds – does *not* alter the set of most plausible worlds. Given that an internally developed history only involves updates with *believed* propositions, it follows that the set of most plausible worlds is the same at *all* stages of an internally developed history. That is, $Min_{\preceq_{s_0}}(W)$ is identical to that of any state s_i in $h = (s_0, \ldots, s_n)$ (of an internally developed history) – i.e., $Min_{\preceq_{s_i}}(W) = Min_{\preceq_{s_0}}(W)$ for any $i \leq n$. Let us call this '(NCP)', for *No Change in most Plausible worlds.*

Given that the set of most plausible worlds is constant for an internally developed history (NCP), it follows that all beliefs and conditional beliefs, which are based on revisions with believed propositions (see footnote 34), of the agent are also constant at all stages. This suggests that the target claim has to be false. To see this, take an arbitrary s_i, where $i < n$, from an internally developed history $h = (s_0, \ldots, s_n)$, such that (i) $\preceq_{s_{i+1}} = \preceq^{\varphi}_{s_i}$ and (ii) $\langle w, h[i+1] \rangle \Vdash B\psi$. Because we focus on *internally developed* histories, all belief revisions are with believed propositions. So, we can conclude from (i), plus Definition 3 and (NCP), that φ is true in all the most plausible worlds of the agent at all stages of the internally developed history. From (ii), plus Definition 4 and (NCP), it follows that ψ is true in all the most plausible worlds of the agent at all stages of the internally developed history. This means that, $\preceq_{s_1} = \preceq^{\varphi}_{s_0}$ and $\langle w, h[1] \rangle \Vdash B\psi$. So, by Definition 6, $\langle w, s_0 \rangle \Vdash B^{\varphi}\psi$. Thus, the target claim is false.[34]

[33] In the conclusion, I will discuss the reliance of the argument on this particular belief revision policy.

[34] Interestingly, imagining something through internal development (i.e., imagining that you already (hypothetically) believe) does affect 'other' conditional beliefs at the different stages. That is, even though we might not gain beliefs in conditionals such that the antecedent is that which we imagine, we might gain conditional beliefs where the imagined proposition is not part of it. For example, consider a model where there are three worlds, such that $V(p) = \{w_2, w_3\}$, $V(q) = \{w_1, w_3\}$, and $V(r) = \{w_1, w_3\}$ and such that $\preceq_{s_0} = w_1 < w_2 < w_3$. If, in this model, we imagine q (which qualifies as an internal development, given that q is believed at s_0), then we have that at the resulting imaginative stage, s_1, the conditional belief $B^p r$ is true, even though this conditional belief is false in the original state, s_0.

This raises a number of interesting questions. For example, would we want to say that imagining q could justify the belief in a (potentially unrelated) conditional $p \rightarrow r$? Whenever Williamson (2007) or Langland-Hassan (2016) talk about imagination, the imagined proposition always features

This suggests that the internal development of pretense imagination *cannot* provide justification for new beliefs in conditionals. Any conditionals that might be imported back into your actual beliefs were *already* believed, for otherwise your internal development would never result in the consequent given the antecedent.

The above argument, one might worry, seems to assume that indicative conditionals express propositions, where as not everyone might agree with this (e.g., Edgington, 1986; Levi, 1988, 1996; see Leitgeb, 2007 for a clear discussion on these and related views). If you think that conditionals do *not* express propositions, they cannot strictly speaking be believed. As Leitgeb points out, "conditionals [on such a view are] accepted by the agent without being believed" (2007, p. 119). For these theorists, 'beliefs in conditionals' simply *are* conditional beliefs (or degrees of belief in the consequent conditional on the antecedent). However, note that the argument against the epistemic usefulness of the internal development of pretense imagination was phrased completely in terms of conditional beliefs. So, the conclusion holds even for those who think that indicative conditionals do not express propositions.[35]

To sum up, despite the fact that it might *seem* as though our imaginative episode makes us believe certain conditionals, it is not the internal development of the imagination that *provides* the justification for the beliefs in these conditionals.

4.6 Epistemic Usefulness of Intervened Content

When focusing on internally developed histories, we saw that we cannot gain knowledge of any *new* conditionals. This is perhaps unsurprising and it seems that what theorists like Williamson (2007) and Langland-Hassan (2016) have in mind is that pretense imagination really comes into its own when we *intervene* some content and then look at the resulting hypothetical belief revisions. As we will see, it is indeed the case that we can come to gain conditional beliefs that we did not have at the root stage, when we have actively intervened content in the imaginative episode.

To show this, let's construct a model where there is a conditional belief at a stage of the imaginative episode and that conditional belief is *not* true at s_0 – i.e., it is a *new* conditional belief. Consider a model such that $W = \{w_1, w_2, w_3\}$, $V(p) = \{w_2, w_3\}$, $V(q) = \{w_1, w_3\}$, and the plausibility orderings per stage are as represented in Figure 4.2 – i.e., $\preceq_{s_{12}} = \preceq^p_{s_0}$ and $\preceq_{s_{21}} = \preceq^q_{s_{12}}$ (only part of the model

as the antecedent of the corresponding conditional. The same holds for the literature surrounding the Ramsey-test and the epistemology of indicative conditionals: we (hypothetically) update our beliefs *with the antecedent* in order to see if the consequent holds. It seems to me not straightforward to defend the position that imagining q *justifies* accepting the new belief in the conditional $p \rightarrow r$, however, more needs to be said about this. Unfortunately, this is outside the scope of this chapter.

[35] Additionally, one might worry that these findings and arguments completely undermine the use of the Ramsey test. However, as we will see below, if we use *intervened* content and then hypothetically update our beliefs (i.e., let the imagination internally develop), then we do get justification for beliefs in conditionals.

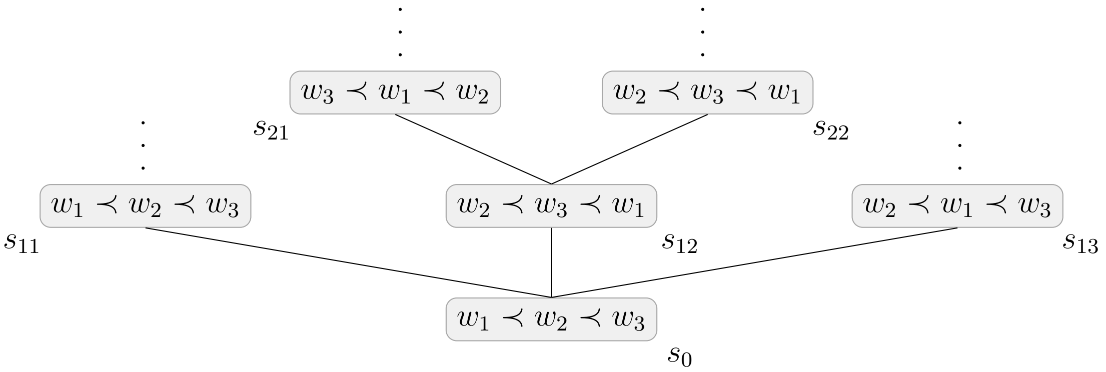

$w_3 \prec w_1 \prec w_2$
s_{21}
$w_2 \prec w_3 \prec w_1$
s_{22}
$w_1 \prec w_2 \prec w_3$
s_{11}
$w_2 \prec w_3 \prec w_1$
s_{12}
$w_2 \prec w_1 \prec w_3$
s_{13}
$w_1 \prec w_2 \prec w_3$
s_0

Figure 4.2: *New conditional beliefs from intervened content.*

is represented). We take the actual history to be $h = (s_0, s_{12}, s_{21})$. Note that both developments are intervened content, as we assume that the explicit input is also intentionally added. For our argument, we focus on the second intervention (i.e., the one *within* the imaginative episode, not the one that starts it).[36] After the second upgrade with q, the agent hypothetically believes p; that is, at stage s_{12} the agents has a conditional belief: they believe p conditional on q. However, it is easy to see that this is *not* the case at the initial stage: $\langle w, s_0 \rangle \not\Vdash B^q p$. So, it seems that we are able to gain *new* beliefs in conditionals by upgrading our (hypothetical) beliefs with intervened content. In our toy example, we gain justification for the belief in $q \rightarrow p$, which we didn't believe before we engaged in the imaginative episode (i.e., at s_0).[37]

4.7 Pretense Imagination and the Epistemology of Possibility

Allowing the intervened content to internally develop seems to get us new conditional beliefs, which, according to Langland-Hassan (2016) and Williamson (2016a), are justified on the basis of this imaginative exercise. The question that arises, however, is what good this does us as epistemologists of possibility, for it seems that there are virtually no constraints on what we use as intervened content: we can intervene

[36] One could also construct a model where the imaginative episode starts with internally developed content and still make the same argument. In such a case, the model would be as above, but with $\preceq_{s_{14}} = \preceq^q_{s_0}$ and $\preceq_{s_{22}} = \preceq^p_{s_{14}}$.

[37] This toy model is of course a simplification and there are probably a number of internally developed steps in between (which is what, e.g., Williamson seems to mean with 'develop the supposition'). However, as we saw with internally developed histories, the set of most plausible worlds after the last intervened upgrade is the same throughout the following internally developed upgrades. So, for simplicity, we ignore these potential intermediate internally developed upgrades.

content that is true, false, impossible, et cetera.[38] Intervened content is just content transferred from the intention (to imagine something) to actually imagining it.[39] Correspondingly, intervening content does not, by itself, carry any epistemological weight; it is the mental equivalent of handing yourself a dollar *to further the internal development* (paraphrased from Langland-Hassan, 2016, p. 61).

Consider how this affects potential epistemologies of possibility based on such conditionals (e.g., Williamson, 2007). When considering the epistemological role of conditionals in the epistemology of possibility, we see that researchers often focus on providing us with justification for believing the possibility of the consequent.[40] We saw, in Section 4.5, that if the input is already believed, we gain no *new* conditional beliefs. So, in order for pretense imagination to be epistemically useful, we should not believe the intervened content to be true (either because we believe it to be false or because we are agnostic about its truth-value). Even though an imaginative episode with an intentionally intervened (believed to be) false proposition might result in justification for a belief in a conditional, how does this help us in determining the modal status of the consequent? In particular, given that we do not believe the antecedent to be true, it might be that the antecedent is 'merely actually false' (i.e., false in the actual world, though possibly true) or 'necessarily false' (i.e., impossible). Consider the following pairs of conditionals to see this:

(6) If Amy squared the circle, she becomes a famous logician. (Ripley, 2012)

(7) If Tom works in his office, he is sitting in a comfortable chair.

Let's assume that we justifiably believe both conditionals based on our pretense imagination and we believe both antecedents to be false. If we are unaware of the modal status of the antecedents, what good does knowledge of these conditionals do us in the epistemology of possibility? So, the crucial issue is how we determine that the hypothetical situation (i.e., the antecedent) is possible. Once we have *indepen-*

[38] Note that in the model discussed in this chapter, we can only imagine 'conjoined' impossibilities. That is, if we upgraded our simulated beliefs with φ and at some later point with $\neg\varphi$, we can be said to have imagined φ and, in the same episode, $\neg\varphi$. However, we cannot imagine 'atomic' impossibilities in this model (e.g., 'unicorns exist,' 'there is a round square,' etc.). A more faithful modelling of pretense imagination should ultimately allow for these, potentially with additional impossible worlds (see Berto, 2017). The fact that the model does not allow for imagining impossibilities is a shortcoming of the model, which can be fixed by, e.g., the incorporation of impossible worlds, and *not* a flaw in the argument. As discussed in Chapter 2 (Sections 2.3 and 2.4), it is very plausible we *can* imagine impossibilities.

[39] Remember that this often goes unconsciously, so it may not feel like you *intend* to imagine these things (see footnote 11).

[40] This is not a surprise. We just saw that the antecedent – i.e., the intervened content – can be anything and that simply being a supposed proposition does not carry any epistemological weight. Furthermore, if it is true that we end up believing the corresponding conditional, then we believe it to be actually true. So, the possibility of the conditional would be of the 'uninteresting' kind of knowledge of possibilities resulting from the actuality principle (see Chapter 1).

dent evidence for the possibility of the input proposition, we might use indicative conditionals (and the corresponding imaginative exercises) to *extend* our knowledge or beliefs. But, prior knowledge of the modal status of the antecedent is crucial; without it pretense imagination is of no help in the epistemology of possibility.

Considering the way most conditional-based epistemologies of possibility work, we can see that, in general, they rely on the *transfer of possibility* from the antecedent to the consequent. That is, if the antecedent is known to be possible and the conditional is believed to be true, then you can believe that the consequent is also possible. For example, Williamson (2007, p. 156) argues that he relies on the counterfactual conditional, *because* counterfactual conditionals satisfy the principle of possibility:

$$\text{(Possibility)} \qquad \varphi \mathrel{\Box\!\!\rightarrow} \psi \vDash \Diamond\varphi \rightarrow \Diamond\psi$$

Similarly, Kment (2006, 2014) argues that we need to know which "grade of possibility" is "attached" to the antecedent, for only then do we learn in which sphere of worlds around actuality the consequent holds (2014, p. 4).[41] This means that in order for the beliefs in the conditional to be useful as a tool for the epistemology of possibility, we need to be able to know what the modal status of the antecedent is.

Prior Modal Knowledge

Williamson (2007, ch. 5) seems to suggest that (something like) pretense imagination is crucial for his conditional-based epistemology of modality. What this chapter shows is that it can only play an *extending*-role. In order for such a condition-based epistemology of possibility to come off the ground, we need prior knowledge of the modal status of the antecedent and pretense imagination does not seem to be able to provide this.

These theorists might well be right that the imagination involved in the epistemology of the relevant conditionals is such that if the input is possible, the resulting conditionals will have possible consequences. Yet this leaves the crucial question of how we should determine the modal status of the antecedent itself. As Gregory (2017) puts it,[42]

> while the described method may well produce beliefs about possibility that tend to be right, our justification for holding that it does so depends upon our being entitled to assume the customary possibility of

[41] This is crucial for Kment as he, as opposed to Williamson, thinks that there are spheres that include impossibilities. That is, Kment rejects vacuïsm with respect to counterpossibles, whereas Williamson (2018) accepts it. See Berto et al. (2018) for a discussion of vacuïsm and counterpossibles.

[42] Gregory (2017) argues against Williamson's epistemology more generally. For example, he argues that Williamson's epistemology of modality fails to work as it is not obvious that our ordinary capacity to evaluate Williamson's conditionals are reliable when it comes to the cases relevant for the modal implications of such conditionals. See also Jenkins (2008); Roca-Royes (2011b); and Tahko (2012) for critical discussions of Williamson's (2007) epistemology of modality.

> the propositions that serve as the starting-points of applications of the relevant process. (p. 834)

The moral of this story is that if we want to use conditionals to gain knowledge of possibilities, we need *prior modal knowledge.*

To sum up, it seems right that pretense imagination can provide us with justification for beliefs in indicative conditionals (as Langland-Hassan, 2016 and Williamson, 2016a suggest). However, it is not the case that this does any work within the epistemology of possibility. The use of such conditionals might *expand* our modal knowledge, but it relies on having *prior* knowledge of the modal status of the antecedent. This means that pretense imagination cannot be the foundational method for determining whether something is possible. As with QALC imagination discussed in the previous chapter, leaving this prior modal knowledge unexplained results in an unsatisfactory epistemology of possibility; one that fails to address the central question of the field: "[h]ow can we come to know (be justified in believing or understand) what is possible" (Vaidya, 2016, §0).

4.8 Conclusion: Potential Objections

This concludes the evaluation of pretense imagination and its role in the epistemology of conditionals and the epistemology of possibility. With regards to the former, Langland-Hassan (2016), Williamson (2016a), and others are right in thinking that pretense imagination seems to be able to justify new beliefs in conditionals. This happens when we forcefully intervene content and then allow it to internally develop. However, this kind of conditional reasoning *cannot* play a fundamental role in the epistemology of possibility: it might be used to expand our modal knowledge, though this requires prior knowledge of possibilities. This prior knowledge of possibilities can itself *not* be justified through pretense imagination.

In this conclusion I want to discuss a number of objections to various parts of the epistemological discussions. I will discuss (i) the wrong formalism objection; (ii) the actuality worry; (iii) the wrong conditional objection; and finally (iv) the wrong imagination objection, in turn.

The Wrong Formalism Objection

One might worry that the reason why the internal development is not epistemically useful is because of the particular, idealised, formalism in which I chose to model pretense imagination. Perhaps wrong choices were made and the conclusions would be different if one were to use a different formalism.

In response, note that all we relied on from the formalism is the fact that revising one's beliefs with something that is currently believed does *not* change the set of most plausible worlds. This seems like a plausible assumption and is not a particularity of the formalism used here. That is, the arguments of this chapter hold for any belief revision policy that is such that updating one's beliefs with a proposition that is believed does not change the set of most plausible worlds.[43] In particular, when we think about the project of this chapter – i.e., modelling pretense imagination – this seems like a very plausible assumption. If in an imaginative episode you update your (hypothetical) beliefs (of that particular state) with something that you believe (at that particular state), you do not all of a sudden change your (hypothetical) beliefs; nothing really changes.

Additionally, the argument relied on a very sensible epistemology of conditionals, that linked conditional beliefs with beliefs in conditionals. Although there are some logical results that affect the *logical* equivalence between these things, the epistemological and psychological relation that we relied on is supported by empirical data and is independent of the formalism used.

The Actuality Worry

The reason why we concluded that pretense imagination cannot provide a satisfactory epistemology of possibility is that it requires prior modal knowledge: knowledge that the antecedent is possible. One might respond as follows: if we use only propositions that we believe to be true as antecedents of the conditional, can we then not expand our knowledge of possibilities on the basis of this? For, as discussed in Chapter 1 (Section 1.4.2), whatever is actually the case is possible, so having the initial input believed to be true means that we believe it to be possible as well.

Note that if we use 'believed to be actual'-propositions as antecedents, then we would have to (hypothetically) revise our beliefs with a *believed* proposition. But, as we saw when discussing the epistemological value of the internal development (Section 4.5), this does *not* result in new conditional beliefs. Phrased differently, the only way in which pretense imagination is epistemically useful, is if we do not believe the antecedent of the conditional in question to be true (either because we are agnostic about it, or because we believe it to be false). Thus, we cannot use the actuality principle in combination with pretense imagination to expand our modal knowledge based on propositions that are believed to be actually true.

The Wrong Conditional Objection

Throughout the discussion of pretense imagination providing justification for newly believed conditionals, we have focused on *indicative* conditionals. However, as I've explicitly mentioned a number of times, many who think that conditionals are involved in the epistemology of possibility rely on *counterfactual* conditionals. For

[43] Note that the arguments *do not* require that the plausibility order stays the same when revising our beliefs with a believed proposition. All that we need is that the *most plausible* worlds do not change – i.e., that $Min_{\preceq_{s_n}}(W)$ stays the same.

example, Williamson (2005, 2007) seems to suggest that the epistemology of modality is a special case of the epistemology of counterfactuals and Kment (2006, 2014) argues for analysing modality based on something akin to similarity-spheres of counterfactuals. The worry is that the result that indicative conditionals cannot play a fundamental role in the epistemology of modality is neither here nor there.[44]

Williamson (2016a, p. 118) is rather explicit in that he thinks that the cognitive capacities that underlie the justification of counterfactual and indicative conditionals are largely similar. Of course, he also acknowledges that there must be some difference between the two, due to the difference in truth-value of famous pairs of such conditionals, but he never elaborates on what this difference is supposed to be. The way that Williamson talks about it makes it seem that the difference is insignificant to the epistemology of modality. The arguments here suggest that either this is not so (that is, pretense imagination as modelled here does not (solely) play a role in the epistemology of *counterfactual* conditionals), or, if it is, the use of pretense imagination in the evaluation of counterfactual conditionals that feature in the epistemology of possibility also require problematic prior modal knowledge. In general, the main argument against the use of pretense imagination in the epistemology of possibility concerns the *problematic prior modal knowledge* required. This holds for *any* conditional for which pretense imagination plays a crucial role in its epistemology. For example, even though Williamson's epistemology of possibility relies on *counterfactual* conditionals, rather than indicative conditions, it crucially relies on pretense imagination. The arguments of the chapter affect any conditional for which the epistemology is taken to be one of hypothetical belief revision.

The Wrong Imagination Objection

Another question that might be raised is whether pretense imagination is the kind of imagination that people take to be used in the epistemology of possibility. This is, again, a fair worry. The first thing to note is that pretense imagination is definitely a real kind of imagination and it is very likely that the best way to model it is through hypothetical belief revision (Currie & Ravenscroft, 2002; Nichols & Stich, 2003; Langland-Hassan, 2016). Though, I should stress that Williamson (2007, 2016a) does not explicitly claim that imagination is *exclusively* the recreation of rational belief revision. Langland-Hassan (2016) does talk about simulated belief revision extensively, however, he also notes that there are other mental faculties that imagination might simulate and talks about perceptual simulation in tandem with belief revision.[45] So, potentially there are other mental faculties that imagination might simulate (e.g., Balcerak Jackson, 2018; Gregory, 2019) or imagination might

[44] It seems that in more recent work, Williamson (2016a) *is* talking about indicative conditionals.

[45] Theories such as those of Currie & Ravenscroft (2002) and Nichols & Stich (2003) focus exclusively on imagination as simulated belief revision, as does Langland-Hassan, 2012 in a sense (although even people such as Currie and Ravenscroft allow for additional forms of imagination such as, e.g., desire-like imaginings).

be better understood as representing alternative situations (e.g., Kung, 2010).

I have argued in the previous chapter that the latter kind of imagination seems unable to justify our beliefs in possibility claims. This leaves the option of a recreativist account of imagination that focuses on the recreation of other cognitive faculties than belief revision. In the next chapter I will discuss such theories – i.e., the *appearance-based* approaches (Balcerak Jackson, 2018; Gregory, 2019). These accounts suggest that imagination is the recreation of *perceptual states*. As we will see, these theories have some issues, but it seems that they are *not* susceptible to the objections raised in this chapter – i.e., they do not rely on problematic prior modal knowledge.

Chapter 5
Putting Knowledge from Imagination on Firmer Grounds

> *[When] the psychological systems are being used outside their natural domain [. . .] there's less reason to think that they will be successful guides in [such] foreign terrain*
> – Nichols, 2006a

So far, we have seen that both a purely representational view of imagination and imagination as simulated belief revision fail to account for our ability to gain knowledge of non-actual possibilities. In both instances, we require some form of problematic *prior modal* knowledge. In this chapter, I will turn to another recreativist view of imagination: imagination as *sensori-motor simulation.* I will first discuss a contemporary version of this, the *appearance-based* approach (Section 5.1). I will argue that there are two worries for these accounts: the problem of imaginative blocks and the problem of modal objectivity. After this, I will propose a novel, closely related, theory: *embodied imagination* (Section 5.3). I will argue that the embodied approach does not fall victim to the two problems for the appearance-based approach (Section 5.4). Finally, I will discuss the consequences of an embodied imagination approach to imagination-based epistemologies of possibility (Section 5.5).

5.1 Appearance-based Imagination

For Balcerak Jackson (2018), appearance-based theories of imagination – and recreative approaches in general – are mainly motivated by trying to overcome the *Up-To-Us Challenge*.[1] In this challenge, she pits this epistemic usefulness of imagination against the *voluntariness* with which we seem to be able to imagine things. The worry is that "[i]t is just because forming images is a voluntary activity that it does not instruct us about the external world" (Wittgenstein, 1967, §627). In the previous chapter we already saw that if we focus on *only* the intentional part of imagination, it's epistemological value is nothing more than "the mental equivalent of handing yourself a dollar" (Langland-Hassan, 2016, p. 61).[2] The fact that because of imagination's voluntary nature it "cannot teach us about anything, or at least not about anything that we didn't already know," is what Balcerak Jackson (2018, p. 212) calls the Up-To-Us Challenge.

In order to fully appreciate this challenge, Balcerak Jackson distinguishes between two ways in which imagination might be under our voluntary control: (i) imaginings are mental states that we only engage in when we intentionally choose to do so and (ii) "imaginings are mental states whose content is determined by what we choose to imagine" (2018, p. 212). The former understanding seems false: if you think that daydreaming or instances of mind-wandering are cases of imagination, then they constitute counterexamples to this first claim. It is the second understanding of the voluntariness of the imagination that is worrisome. First of all, the voluntariness makes for a dis-analogy with perception: you might think that perception provides us with justification for our beliefs *because* the world imposes itself on us, we have no voluntary control over our perceptual experiences. Though, as Balcerak Jackson admits, such an "argument by dis-analogy cannot establish the strong conclusion that imaginings cannot provide justification at all" (2018, p. 214). The stronger argument is based on the fact that the voluntariness makes it so that there are no limits to what we can imagine – i.e., imagination is *unrestricted*. If imagination is limitless, then it cannot provide us with justification. The reasoning goes as follows: justification involves ruling out alternative hypotheses; but if imag-

[1] Balcerak Jackson talks about 'recreativist' imagination more generally. However, in order to distinguish the kind of imagination I am interested in in this chapter from imagination as recreating rational belief revision, I use the phrase 'appearance-based' for the recreation of sensory experiences (this is inspired by Gregory, 2010).

[2] This issue is nicely discussed by Langland-Hassan (2016) when he discusses the *Only Top Down* approach to imagination. The idea on this approach, which is not likely to be actually held by people as Langland-Hassan rightly points out, is that "the content of each [imagined] proposition is determined by an intention to imagine a proposition with that very content" (2016, p. 65). On this view, imagination is completely under our voluntary control in that we know exactly what we can and will imagine. Note, though, that this kind of imagination seems to be *epistemically vacuous*, for, as Langland-Hassan nicely puts it, on this view "[i]magination becomes a kind of internal transfer of contents" (idem, p. 61) and "one ends up where one began, epistemically speaking" (idem, 65).

ination is unrestricted, then imagining something does not rule out *any* hypotheses whatsoever. The conclusion of this second, stronger argument is "not merely that it is mysterious how imagination could serve as a source of justification, but rather that the fact that imaginings are up to us makes it impossible for them to provide justification" (Balcerak Jackson, 2018, p. 214).

The recreative conception of imagination, which has been thoroughly developed and defended in cognitive science and the philosophy thereof – most prominently by Currie & Ravenscroft (2002) and Goldman (2006) and more recently by Langland-Hassan (2016); Balcerak Jackson (2018); and Gregory (2019) – offers a way out of this puzzle. According to these views, imagination can mimic other cognitive faculties. That is, imagination is an *offline* version of certain online perceptual or motor counterparts. In this chapter, we focus in particular on the *appearance-based* recreative theories of imagination, which hold that imagination recreates (or simulates) most notably the perceptual faculties, *without* the corresponding sensory input or behavioural reaction to it.[3] So, according to such appearance-based theorists, imagery is, very roughly, the offline counterpart of vision.

Balcerak Jackson (2018) notes that there are two crucial aspects in which imagination *recreates* the (perceptual) process of which it is supposed to be an offline counterpart. Firstly, there is the phenomenological aspect. Imaginings have a particular phenomenological character that is similar to the phenomenological character of the corresponding perceptual experience. For example, when I imagine listening to my favourite song, I imagine *what it is like* to listen to that song; there is a phenomenal experience of imagining listening to that song, just as I would have (though perhaps less vividly) when I would actually listen to it. Secondly, and more importantly, imaginings replicate "the *representational content* of their possible counterparts without actually being perfect copies of those counterparts" (Balcerak Jackson, 2018, p. 218, emphasis added). Remember from our discussion in Chapter 2 (Section 2.2.1) this feature carries a lot of weight in the justificatory role of imagination.

Given that the appearance-based theories of imagination focus on *imagery* (Balcerak Jackson, 2018; Gregory, 2019), these two ways in which imagination mimics perceptions give rise to two further important features of imagination on these accounts. First, these kinds of accounts focus on *objectual imagination* rather than *propositional imagination* (see Chapter 2 and Yablo, 1993). An objectual imagining is, for example, the imagining of a cow, whereas a propositional imagining would be

[3] This is supposed to be opposed to *pretense imagination*. Pretense imagination is also a recreative theory of imagination, but focuses on the recreation of our rational belief revision faculties. Pretense imagination and appearance-based imagination are both merely a *subclass* of recreativist accounts of imagination. Ultimately, it is likely that a recreativist account of imagination involves a subtle mixture of both objectual (appearance-based) and propositional (pretense) imaginings and that people such as Langland-Hassan (2016); Williamson (2016a); Balcerak Jackson (2018); and Gregory (2010, 2019) have something like this in mind.

the imagining that there is a cow. Remember that objectual imaginings are "mental states in which a subject bears an imagination relation to an object or an event [...] rather than to a proposition" (Balcerak Jackson, 2018, p. 201). Secondly, on appearance-based accounts, imagination is supposed to provide *phenomenal evidence* as opposed to *physical evidence*. That is, to the extent that imagination gives us evidence or justification for something, it provides us with evidence about possible experiences, as opposed to evidence that the world is or could be a certain way. As Gregory puts it, "the sensations which we imagine having when engaging in sensory imaginings seem to be the kind of *sensations that we could have*" (2010, p. 336, emphasis added).

The recreative nature of imagination nicely explains the way in which appearance-based imagination gets its justificatory status and gets out of the Up-To-Us challenge. As this is important, let me explain it in a bit more detail. In particular, let me stress how this relates to the constraints on imagination that most researchers agree are needed to give imagination its epistemological impact (see Kind, 2016a; Kind & Kung, 2016a). The puzzle of imaginative use suggested that in order for imagination to be epistemically useful, while also allowing us to dream up the most fantastical situations, it has to be restricted.[4]

Balcerak Jackson describes how it is that appearance-based imagination is restricted due to its recreativist nature as follows:

> [T]he idea is not merely that imaginings justify us in beliefs about how things could look because their content and phenomenal character *resembles* the content and the phenomenal character of perceptual experiences. Rather, imaginings play this role in virtue of being by their very nature *derived from* or parasitic on perceptual experience, which in turn informs us about the visible properties of objects. It is because imagination is constitutively a capacity to recreate perceptual experiences [...] that it can tell us how things look. (2018, pp. 221-222, original emphases)

Balcerak Jackson acknowledges that imagination needs to be constrained in order to be epistemically useful and points out that it is not the fact that imaginative episodes resemble the experiential content of perception that enables imagination to play an epistemic role. It is that imagination by its very nature *derives from* perceptual experiences, that imagination can provide justification. I think it is fair to say that what Balcerak Jackson has in mind is that this, in turn, is because perceptual experiences are constrained by the 'make-up' of our perceptual machinery. So, I suggest that the reasoning is as follows: our perceptual mechanisms are inherently constrained (by the physiological nature of our eyes, neurons, etc.), thus,

[4] The puzzle of imaginative use is the tension that seems to exist between the fact that imagination allows us to explore the most fantastical situations, yet also seems to be able to provide us with new knowledge. The puzzle is elaborately discussed in Chapter 2 (Section 2.2).

if imagination is parasitic on perception, it will be similarly constrained. It is in this way that imagination is constrained and that it gets its justificatory force.

With this explanation of what an appearance-based account of imagination is and how it grounds the justificatory role of imagination, it is also easy to see how it answers the Up-To-Us Challenge. "Despite its voluntary nature, imagining can provide us with justification because what we imagine is *constrained* by the recreative nature of imagination" (Balcerak Jackson, 2018, p. 222, emphasis added). Moreover, one might argue that, on such accounts, only the explicit start of an imaginative episode is voluntary, but the development of such an offline simulation happens, to a certain extent, involuntarily (Langland-Hassan, 2016; Williamson, 2016a; Balcerak Jackson, 2018).

5.2 Imaginative Blocks and Objective Evidence

There are two issues that are troubling for the current appearance-based theories of imagination, in particular with respect to imagination's role in the epistemology of possibility. The problem of imaginative blocks and the problem of objective evidence. I will discuss these in more detail in turn.

Let me stress that the issues raised with regards to appearance-based theories of imagination are significantly different from those raised against QALC imagination and pretense imagination in the previous two chapters. There, the main issue was that the accounts required prior knowledge of necessities and possibilities, respectively, in order to provide a satisfactory epistemology of possibility. The issues raised for the appearance-based theories, on the other hand, involve the *kind* of modal knowledge imagination justifies (this is the target of the problem of objective evidence) and what it is that *grounds* imaginations justificatory force (this is the target of the problem of imaginative blocks).

5.2.1 The Problem of Imaginative Blocks

The problem of imaginative blocks, as we will see below, comes from the interaction of imaginative blocks and *unchosen* constraints. Throughout this part of the dissertation, I noted that imagination has to be constrained in order for it to be epistemically useful. However, an important subtlety concerning the constraints on imagination is rarely made in the literature. When Kind and Kung initially discuss the issue of constraints on imagination, they point out that these constraints can be "*either* architectural constraints *or* constraints that we can willingly impose" (2016a, p. 2, emphases added). This distinction, though rarely noted, has important consequences. So, let us spell out the difference between these two in a bit more detail. In one sense, we are able, as imaginers, to *place* constraints upon our own imagination. Knowing that we are in the business of trying to acquire knowledge, as opposed to engaging in mere flights of fancy, we can *choose* to constrain our imagination so that it functions to fulfil the particular task at hand. For example, by restricting imagination to be reality-oriented (Kind, 2016a; Williamson, 2016a). These are the kinds of constraints that most researchers seem to focus on in relation to the epistemic usefulness of imagination (see, e.g., Kung, 2010; Kind, 2016a; Kind & Kung, 2016a; Williamson, 2016a).[5]

This is significantly different from the way in which imagination is restricted on the recreative accounts of imagination – i.e., pretense imagination, the appearance-based theories, as well as the embodied account that I will present below. On such accounts, imagination is restricted not as a matter of choice, but simply as an *inherent feature* of the kinds of imaginers that we are with the kinds of brains and bodies that we happen to have – i.e., imagination is architecturally constrained. These are constraints on our imagination that are not under our control, but due to our neuro-physiological make-up. Our minds are limited in many ways. For example, we cannot imagine every detail of what it would be like to live for a thousand years, as this would presumably take longer than our lifetime (Van Leeuwen, 2013) and most of us cannot imagine episodes with the same richness and detail as conscious perceptual experience.[6] These limitations on our imagination are not a matter of choice, they are just features of the kinds of imaginers we are and the kinds of systems that give rise to imagination (remember the quote from Balcerak Jackson discussed above).[7]

We can use this distinction between chosen and unchosen constraints to raise our first worry for the current appearance-based theories of imagination.

Imaginative Blocks and Introspective Access

The problem of imaginative blocks, roughly, is our inability to judge the epistemic consequences of our failure to imagine something. In particular, I use this problem as an example to raise a worry for the ways in which the epistemic role of imagination is supposed to be grounded according to appearance-based theories of imagination. It is not that the epistemology of possibility based on appearance-based imagination gets the wrong predictions or relies on problematic prior modal knowledge. It is that the *explanation* that they give for the epistemic usefulness of imagination does not seem to work. Let me explain and make this more precise.

[5] Stuart (2019) calls these 'constraints on imagination$_1$', whereas he calls, what we will call 'unchosen constraints', 'constraints on imagination$_2$'.

[6] Though people with *hyperphantasia* might be an exception (see, e.g., Zeman et al., 2020).

[7] Another example comes to mind: we cannot imagine extremely morally repugnant scenarios as the imaginative episodes are too harrowing to sustain (Gendler, 2000; Weatherson, 2004). These cases are known under the label 'imaginative resistance' and concern a *different* kind of unimaginability and we will set them aside as such (see Nichols, 2006a, p. 246, fn. 16 for similar remarks). Though see Kim et al. (2019), who argue for a kind of imaginative resistance different from that discussed by Gendler and which might be a more relevant form of imaginative resistance. The kind of imaginative resistance that we are setting aside is the former kind.

The problem for appearance-based theories comes from the phenomenon of *imaginative blocks*. You experience an imaginative block when there is something that you simply *fail* to imagine or, more tersely, that you *currently cannot* imagine. It is important to stress that imaginative blocks need not involve the *impossibility* of imagining something (whatever that may be). It simply means that currently, with the current constraints on imagination or within your current cognitive state, you cannot (or fail to) imagine something. The *problem* of imaginative blocks arises due to the interaction of imaginative blocks and *unchosen* constraints.[8] Consider what could be the cause of an imaginative block when there are constraints in play (whether chosen or unchosen): (i) we may fail to imagine something due to the constraints on imagination or (ii) we may fail to imagine something because what we are trying to imagine simply cannot be the case.[9]

The problem of imaginative blocks is an epistemic problem: when you fail to imagine something and there are unchosen constraints in play, you may not be in a position to determine whether the source of the imaginative block is (i) or (ii). When we choose our constraints, this is not so problematic, as the following case shows. We imagine how to prepare for a trip from St. Andrews to Amsterdam and, in order for this imagining to be epistemically useful, we constrain our imagination to be reality-oriented. By constraining our imagination thusly, we fail to imagine that our trip takes less than 30 minutes. However, it is not impossible to travel 1300km/h and we should realise that our failure to imagine this is due to our chosen constraints. Now, consider when we fail to imagine something and there are unchosen constraints at play. In this case, because the constraints are unchosen, there is no guarantee that we are aware of these constraints nor that we can introspectively access them. For consider an example of the unchosen constraints on an appearance-based account: the bounds of the neurological make-up of our perceptual system. It is unlikely that we would ever be in a position to access these constraints by introspectively reflecting on our imaginative seemings. This then gives rise to the problem of imaginative blocks. We are not always (if ever) in a position to be aware of unchosen constraints on imagination 'from the inside'.[10]

[8] The problem of imaginative blocks was initially raised by Blackburn (1993). However, the problem as formulated by Blackburn is different from the way that it will be understood here. For Blackburn, the problem involved our ability 'to make something' of the thing we could not imagine. See also Nichols (2006a) for a discussion of Blackburn's problem of imaginative blocks.

[9] It is important here that I am being deliberately vague with 'it cannot be.' This is meant to capture that for whatever modality you think imagination plays a role in its epistemology, you cannot imagine that which is impossible according to *that* modality.

[10] Note that this is not to say we might not be aware of *other* limitations of my imagination. For example, when I imagine robots in a particular way (e.g., that they can speak) and 'fail' to imagine certain other characteristics of robots (e.g., that they have a human-like physique), I might very well realise that this imagining is shaped and limited by my recently watching *2001: A Space Odyssey*. But these are not the kind of limitations I mean when I say we are unaware of them. I mean to say that we are (introspectively) unable to access the limitations that are the result of our cognitive machinery. Thanks to Deb Marber for raising the Space Odyssey example.

Of course, appearance-based imagination supporters do not suggest that unimaginability leads to knowledge of impossibility and mostly only focus on the link between imaginability and possibility. So, it might seem that they need not be fazed by this. However, all that the problem of imaginative blocks is supposed to show is that there are *limitations to the introspective access* we have to our imagination and its features (e.g., the constraints thereon). Crucially, the problem relies on the fact that we are not *able to* introspectively determine why we have the imaginative blocks we have. It is not (only) that we are poor judges of our own imagination (e.g., Richman et al., 1979; Mitchell & Richman, 1980; Intons-Peterson, 1983; Goldston et al., 1985), it is that *we might be fundamentally unable to introspectively assess the effects of such unchosen constraints.* Strengthening our introspective skills will not help.

Note that if we are principally unable to access features of our imagination introspectively, then this seriously undermines most *internalist* epistemologies of imagination. Remember from Chapter 1 (Section 1.5.1) that we are interested in how it is that imagination justifies certain beliefs or, more tersely, what is it that grounds the justificatory role of imagination. These findings suggest that it *cannot* be any reference to anything introspectively accessible, for the problem of imaginative blocks shows that we might *principally* be unable to access such internal grounds through introspection. Thus, the problem of imaginative blocks shows that internalist perspectives on the epistemology of imagination are untenable: we might be principally unable to access whatever it is that grounds the justificatory force of the imagination.[11]

Remember that, following Stuart's (2019) description of a process-based epistemology – i.e., "what is it that enables [imagination] to obtain certain epistemological properties" (p. 4) – we are interested in the question of what it is that grounds the epistemological properties (in our case, its justificatory role) of imagination. With respect to this question, part of what is doing the work are the inherited restrictions on imagination, as we saw above. However, appearance-based theorists also often appeal to *introspection* (though I am not claiming that Balcerak Jackson (2018) and Gregory (2019) explicitly express a preference for an internalist epistemology of imagination).[12] Consider for example these quotes from the two most prominent advocates:

> [Imagination] will only enable us to justify some beliefs about possibility if imaginings sometimes produce appearances of possibility. Do they? The obvious way of tackling that question *is to look and see—to examine*

[11] Remember that we understand internalism as the view that whatever grounds the justification of one's beliefs is internal and, more importantly, accessible to the agent (see Chapter 1; see also BonJour, 2003; Greco, 2014; Pappas, 2017). So, the problem raised here seems to undermine even the weak version of access internalism (see Pappas, 2017, §2).

[12] In general, discussions in the epistemology of modality or in the epistemology of imagination are rarely phrased explicitly in these 'more traditional' epistemological debates.

> *the introspective evidence.* (Gregory, 2010, pp. 327-328, emphasis added)

> Going through this exercise thus plausibly gives one prima facie justification for the general belief that one's perceptual experience does not permit one to experience two colours as co-located. [Why? Because] [o]ne will quickly *notice* that it cannot be done [as] one *uses one's own mind as the experimental lab* for this toy study.
>
> (paraphrased from Balcerak Jackson, 2018, p. 223, emphasis added)

In both cases the authors justify the justificatory role of imagination by appeal to introspection (either explicitly or by reference to what one 'notices' in 'one's own mind'). In these instances of the appearance-based approach, what gives the imagination its justificatory power are the 'seemings' that these imaginings produce, which are wholly internal and introspectively accessible to us.[13] Whether or not Balcerak Jackson and Gregory intend to be in line with a roughly internalist epistemology, the problem of imaginative blocks shows that we might not be able to introspectively determine if our imaginative episodes are so constrained to be epistemically useful. Even if it is the case that actual imaginative blocks are not very frequent, the issue raised here is not affected. The problem of imaginative blocks shows that whatever it is that grounds the epistemic usefulness of imagination should not be dependent on introspection.

So, to summarise, the problem of imaginative blocks, for our purposes, is the fact that appearance-based theories of imagination in the epistemology of possibility *cannot* appeal to our introspective seemings as that what is supposed to ground the justificatory role of imagination. This conclusion stands *regardless* of whether Balcerak Jackson (2018) or Gregory (2019) in fact holds something like that or whether we frequently experience such imaginative blocks.

5.2.2 The Problem of Modal Objectivity

The appeal to introspectively accessible seemings is perhaps explained by the fact that the appearance-based approaches of imagination initially focus on *phenomenal evidence* – i.e., evidence of possible *sensations* one could have. However, when we engage in the epistemology of modality, in which imagination plays a significant role, we are interested in knowledge of *objective modal facts* (see, e.g., Williamson, 2016b; Strohminger & Yli-Vakkuri, 2018a).

Balcerak Jackson (2018) and Gregory (2019) both provide an argument linking this phenomenal evidence to the kind of objective modal possibilities that we are

[13] See BonJour (2003, sec. 5) for a full-blown internalist account that is similarly reliant on 'seemings'.

interested in in an epistemology of possibility. For example, Balcerak Jackson puts it as follows:[14]

> I argued that there is no reason to believe that imagination gives us direct insight into metaphysical possibility. However, recreative imagination can perhaps give us indirect insight into metaphysical possibility. Here is a tempting line of reasoning: as we have seen, imagining p gives us prima facie justification for believing that p is a way things could look; but if p is a way things could look, then it could also be the case that things veridically look as if p. And if things could veridically look as if p, then things could be that way, that is, possibly p. (2018, p. 224)

Let's assume that the last premise is uncontroversial: if the way things look is veridical, then we are justified in believing that things are the way they look. The first premise is built into what it is to imagine something on the appearance-based theory of imagination. So the crucial step is the second premise: if things could look a certain way, then they could veridically look that way. What is important for us, is that this second premise seems to rely on some *modal* assumptions. For example, the way that Balcerak Jackson elaborates on what this second premise claims is as follows: "[t]he premise says in effect that, for every *perceptual content*, there is a possible subject that has a perceptual experience with this content and that represents the world as it really is" (2018, p. 224, original emphasis). This seems to implicate that there is a possible world, where there is a subject that has the perceptual experience that you have and in that possible world the perceptual experience of that subject is such that it represents that world as it is. If this is so, then the argument requires an overtly modal premise.

Gregory (2019) gives a more elaborate version of the same line of thought.[15] In the first part of his paper, Gregory presents his argument to the effect that we are *prima facie* justified in taking what is depicted in a visual image to be metaphysically possible. We can reconstruct this argument for the relation between imagery and possibility as follows:

1. Someone has a visual image of p. (p. 5)
2. Having a visual image of p, shows things to look like p. (p. 5)

[14] As Balcerak Jackson (2018, pp. 224-225) points out, the argument from appearances to objective modal knowledge, if it works, is not susceptible to the issues of the Kripke-Putnam cases. The reason for this is that, even though we can entertain the possibility that, e.g., water is not H_2O, we cannot *perceive* this. So, there is no perceptual content representing water being distinct from H_2O.

[15] Note that Gregory (2019) talks about imagery as opposed to imagination, but his notion of imagery can be seen as the simulation of perceptual experiences and is thus a form of what we have been calling appearance-based imagination.

3. How things are shown to be in the image are reliability-compatible. (p. 6)

4. Something is reliability-compatible if there is a possible world where someone has the experience of things looking like p and this experience within the visual reliability conditions. (p. 6)

5. If someone experiences things looking like p under the visual reliability conditions, then things tend to be really like p. (p. 3)

6. So, there is a possible world where things tend to be really like p. (from 1,2,3,4,5)

7. Thus, p is possible. (from 6)

The core of the argument is the joint assumption of (3-5). If we simplify this, by reducing the terminology in all these assumptions, we get the following main assumption:

> The way things are shown to be like in the image are such that there is a possible world where someone has the experience of things looking like p and this experience is such that things tend to be really like p.

It is likely that Gregory's formulation of the argument from the phenomenal evidence of appearance-based imagination to evidence for beliefs in objective modal facts is similar to what Balcerak Jackson has in mind and it might seem to be the best appearance-based theorists can do. However, this argument relies on a very strong assumption with a potentially problematic modal aspect (i.e., 'that there is a possible world where...'). This would mean that in order for me to gain justification for a belief in a possibility claim on the basis of imagining it, I would need to have a prior belief (or assumption) that it is possible for someone to have an experience of things looking like the way I imagined them. The question then becomes, how do I know that there it is possible for someone to have such an experience? Moreover, these accounts leave unexplained *why* we should think that our imagination would be reliability-compatible – i.e., why imagery would satisfy the above assumption (Gregory, 2019, p. 5). Either way, the assumption is rather strong.[16] Preferably, we would have an account of imagination that does not need to rely on such an assumption while securing the justification for beliefs in objective modal facts.

With these two worries on the table, we can turn to a novel theory of imagination: embodied imagination. Yet, before we continue, let me stress a terminological

[16] Whether or not the assumption also requires *problematic* prior modal knowledge is something that I leave aside for now. All I need is that the assumption is so strong that it warrants looking for theories that need not rely on (something similar to) it.

issue. The appearance-based theories of imagination discussed above and my proposed theory of embodied imagination (Section 5.3) are closely related *as theories of imagination.* As a theory of imagination, the appearance-based approach suggests that imagination is the recreation of perceptual states, whereas the embodied theory takes imagination to be sensori-motor recreation, which also includes the recreation of perception. However, as a research programme, embodied cognition suggests a completely different way of looking at the mind, perception, and imagination. This, as we will see below, has significant consequences for any sort of epistemology *based on* these theories of imagination. So, in a sense the embodied theory of imagination can be seen as subsuming the appearance-based theory; but only *as theories of imagination.* There is an additional aspect to the embodied theory of imagination, which includes adopting the perspective of embodied cognitive sciences, that cannot be thought of as 'subsuming' or being otherwise related to the appearance-based approach to imagination-based epistemologies of modality. It is a completely different way of looking at things. This latter aspect will have significant effects, for example resulting in a crucially different view of the justificatory role of imagination. This different perspective will allow us to respond to the above worries.

5.3 Embodied Imagination

I propose a new way of looking at the imagination and the cognitive processes that give rise to it: *embodied imagination.*[17] That is, imagination as restricted by the kind of embodied, cognitive agents we are – i.e., looking at imagination from the perspective of embodied cognitive science. On such an account imagination is seen as sensori-motor simulation that may or may not be accompanied by relevant phenomenology and that is shaped by the environment we've developed in.

Let me first explain what I mean by 'embodied cognition' (as this is a rather vague term) and then elaborate on how this restricts imagination.[18] Then, I will turn to the two issues raised for appearance-based theories as well as the limitations of an embodied imagination-based epistemology of possibility.

5.3.1 Varieties of Embodied Cognition

Embodied cognition is often used as an umbrella term for a large number of different approaches, and is thus best characterised as a research programme, rather than a

[17] Clavel-Vázquez & Clavel Vázquez (2018) have, independently of Max Jones and I, looked at the effects of embodiment on imagination. Their work focuses more on the role that embodied imagination plays in empathy.

[18] Several authors have hinted at the idea that our body has a significant influence on our imaginative capacities, e.g., in the context of the discussion of thought experiments (Gooding, 1992, 1994; Fehige & Wiltsche, 2013), but it hasn't yet been developed to the extent that it deserves.

specific theory (Wilson, 2002; Shapiro, 2007; Steiner, 2014; Wilson & Foglia, 2017). Often, people talk of the *4E* approach to cognition: *e*mbodied, *e*mbedded, *e*nacted, and *e*xtended. Here, we focus on the *embodied* part of these accounts. Given the range of different positions that are frequently characterised as embodied cognition, it is important to be clear about precisely which claims we are endorsing here.

On some accounts, embodied cognition is taken to be a metaphysical claim about what constitutes the mind. According to these approaches, embodied cognition is the claim that the mind extends beyond the brain and is, at least in part, constituted by non-neural bodily processes (and perhaps also parts of the environment, e.g. Clark & Chalmers, 1998). I will remain neutral with respect to this *constitution claim.*[19] On other, more radical accounts, embodied cognition involves rejecting appeals to representations or computational inference in explaining the mind (Chemero, 2009). I will refrain from going so far and allow the mind to be explained in terms of embodied *representations* (i.e., in terms of sensori-motor or action representations, rather than 'semantic' or propositional representations).[20]

The more moderate type of embodied cognition that we will focus on is a claim about the *vehicles* of cognition (explained below). The central idea is that so-called 'offline' cognitive processes, such as thought, reasoning, planning, and, importantly, imagination, utilise the same cognitive resources as are involved in 'online' interaction with the environment such as perception, motor control, etc. (Barsalou, 1999; Prinz, 2004; Barsalou, 2008, 2009; Pezzulo, 2011). Cognition involves re-activation of perceptual, motor, and affective systems, rather than distinct, purely cognitive systems dedicated to processing amodal symbols, in particular, the vehicles of higher cognitive processes strongly overlap with the vehicles that support perception, emotion, and motor control. For example, according to Barsalou's grounded cognition approach (Barsalou, 1999, 2008), thinking about horses involves partial reactivation of the perceptual systems that would be activated by encounters with horses, the motor systems that would be activated by interacting with horses, the affective systems associated with one's emotional attachment to horses as well as parts of other sensori-motor systems that have become associated with horses, such as parts of the auditory system associated with hearing the word 'horse' and parts of the motor system associated with saying or writing the word 'horse' (see also Dove, 2014).

On Barsalou's account, there is a tight link between the content of a given offline activation in a system and the online function that the system plays. However, not all embodied accounts need to accept this link.[21] For current concerns, it is

[19] Though there is some evidence that relates imagery and eye movement that might lend support for such a constitutional claim (Spivey & Geng, 2001).

[20] However, much of what is being said here may be open to reinterpretation in anti-representationalist terms by those that are committed to these more radical approaches. For example, the idea that imagination involves reactivation of perceptual and motor mechanisms need not commit one to the idea that these mechanisms operate using representations (either online or offline) (see Hutto & Myin, 2012; Hutto, 2015).

[21] E.g., in some cases, embodied cognition theorists merely claim that cognition involves reuse

important to note that, however tight one takes the link between online and offline activation of systems to be, the mere fact that offline cognitive processes involve reactivation of perceptual and motor systems is sufficient to imply that the nature of offline cognitive processes will be shaped by the kinds of perceptual and motor systems that we possess. These motor systems that we possess will in turn be shaped by the kinds of bodies that we have and the ways that we are thus able to interact with our environment.

5.3.2 Mental Imagery and Embodied Imagination

Given this conception of embodied cognition, it is unsurprising that it has some bearing on our understanding of imagination. In typical cases, people tend to associate imagination with conscious mental imagery. For example, when considering whether one is able to fit a piece of furniture through a doorway, it is common for subjects to report conscious mental imagery of adjusting the position of the sofa and comparing it to the doorway. In such cases, there is an apparent similarity between perceptual and imaginative phenomenology. People report seeing things with their 'mind's eye' when engaging in imaginative activity.

Although largely anecdotal, these kinds of descriptions of the phenomenology of imagination provide *prima facie* support to the embodied cognition perspective, since it can explain the similarity between perceptual and imaginative experience on the basis of the overlap between the underlying mechanisms that support online perception, on the one hand, and imagery-laden imagination, on the other.

The embodied link between perceptual and motor systems and the imagery that is often associated with imagination receives support from extensive empirical investigation into *mental imagery*. In a landmark series of studies, Shepard and colleagues demonstrated that when engaging in mental rotation tasks, subjects took longer to carry out more extensive rotations, suggesting that they were carrying out some form of simulated actual rotation of the imagined object (e.g. Shepard & Metzler, 1971). Similarly, Kosslyn and colleagues found that, when subjects were asked questions about imagined objects, they took longer to answer questions about features that were further apart, suggesting that they were 'scanning' the mental images (Kosslyn, 1973; Kosslyn et al., 1978; Kosslyn, 1980). If the subjects were just retrieving facts about the imagined objects, one wouldn't expect these effects. The fact that the temporal dynamics of these processes involving mental imagery are closely tied to the dynamics of online perceptual and motor exploration suggest that the same systems may be involved in supporting both activities, thereby providing support for an embodied cognition approach.

of perceptual and motor systems, without there needing to be an obvious link between online and offline content (Anderson, 2014, p. 99). For example, mechanisms for directing overt spatial attention "online" may be reused when we think about time or number offline (Jones, 2018).

Even though this seems highly suggestive, the apparent viability of an embodied account of imagery is not, in and of itself, sufficient to support an embodied account of imagination for a number of reasons. Firstly, there is a long-standing dispute about the format of the representations that support mental imagery. While Kosslyn and his supporters favour imagistic representations, others argue that imagery involves symbolic or propositional representations (e.g., Pylyshyn, 2002). As things stand, neither side of the debate should be seen as favouring an embodied account. The embodied cognition approach would predict that the same format will be utilised in perception and action as is utilised in imagination, so taking a view on the format of mental imagery will only undermine an embodied account if the format of mental imagery is argued to be *distinct* from the format of representations utilised in perception and the control of action.

Secondly, the apparent phenomenological similarity between perception and mental imagery may be based on a misconception about the nature of perception. It is common to think of perception as providing us with picture-like snapshots of the world. However, perception is an active and dynamic process, which rarely if ever delivers us with anything like a static image (O'Regan & Noë, 2001; Noë, 2002, 2004). As such, the appearance that imagination provides us with static images may render imagination less, rather than more like, perception. The debate about the underlying format of mental imagery may have been unwarrantedly restricted to considering either static image-like representations or static propositional representations, when neither adequately capture either the active dynamic experience associated with either perception or the real phenomenology of mental imagery (Thomas, 1999, 2018).

Thirdly, it is not clear that conscious mental imagery is an essential feature of imagination (pace Kind, 2001). There seem to be clear cases of imagination that do not involve conscious mental imagery. Moreover, many report severely impoverished or even absent mental imagery in the case of subjects with aphantasia (Zeman et al., 2015, 2016), yet these people are no less capable of engaging in imagination. Conversely, there may be cases of engaging mental imagery that do not qualify as imaginative activities (Nanay, 2010). As such, one should not expect an embodied account of imagination to be solely motivated by considerations about the nature of mental imagery.

Embodied Imagination without Mental Imagery

Despite the fact that the existence of perception-like mental imagery in imagination may lend some support to an embodied approach, embodied cognition does not predict that phenomenology of this kind will always accompany cognition. The embodied cognition approach suggests that all offline, higher cognition involves reactivation of perceptual, affective, and motor systems. However, the majority of our offline cognition is not accompanied by related phenomenology. Many of the more surprising pieces of evidence in support of embodied cognition are surprising precisely *because* they reveal effects that suggest involvement of perceptual, affective, or motor systems despite the absence of accompanying conscious perceptual, emotional, or action imagery (see, e.g., examples discussed in Clark, 1998).

Even though the embodied account of imagination receives support from the fact that it provides a convincing explanation of mental imagery, it is important to note that invoking embodied cognition in explaining imagination by no means commits one to the idea that imagination always or essentially involves mental imagery. One of the key features of an embodied account of the imagination is that it can explain why there may be embodied constraints on imagination independently of there being any phenomenological similarity between perception, emotion, and action, on the one hand, and imagination, on the other.

5.3.3 From Action to Imagination

Rather than basing the embodied approach to imagination on the apparent similarity between perceptual imagery and conscious mental imagery in imagination, I see it as being more fruitful to focus on avoiding introspective analyses and turn to look at accounts of the evolutionary origins of imaginative capacities. When one considers the question of how imaginative capacities emerged in the first place, an embodied account is strongly supported. Despite our intuitive sense of imagination as closely related to perception, leading accounts of the origins of imagination see it as more closely related to our capacity for selection and control of action. In particular, a number of theorists have explicitly developed embodied accounts of the origins of imagination in systems for the selection and control of action (Jeannerod, 1994; Hesslow, 2002; Grush, 2004; Jeannerod, 2006; Pezzulo, 2011, 2017). According to these accounts, our capacity for imagination is closely tied to *anticipatory* mechanisms that play an important role in motor control.

Engaging in effective goal-directed action requires some way of selecting among a range of possible actions and of monitoring the progress of a selected course of action. The brain needs some way of assessing whether the motor system has adjusted the body in the manner desired. One way to do this would be to continually adjust one's movements in response to sensory feedback concerning how the given action is going. However, "the delays in most sensori-motor loops are large, making feedback control too slow for rapid movements" (Wolpert et al., 1995, p. 1880). Thus, rather than waiting for feedback, the brain actively anticipates the outcome of a given action, producing a "forward model" of the expected action dynamics from which the sensory consequences of the action can be predicted (ibid.). The predicted sensory outcome can then be compared with the actual resulting sensory input, correcting for errors to derive a new model of the resultant bodily state. The important aspect of this theory for current concerns is that basic motor control already requires *simulation* of bodily dynamics and their sensory consequences. The basic ingredients of motor control already include an *emulator* of bodily processes (Grush, 2004).

Predicting the sensory outcomes of a range of potential actions can then also play a role in selecting which action to engage in. Importantly, since at any one time one will be faced with a range of mutually incompatible possible actions, some of the simulated sensory consequences will never actually take place. As such, even online motor control requires something akin to imagination, whereby non-actual scenarios are represented, albeit usually unconsciously (Pezzulo, 2011; Burr & Jones, 2016; Clark, 2016).

Once an organism has the capacity to generate forward models for the selection and control of action, they can then run the same process offline in the absence of any actual actions. Organisms can simulate what the sensory consequences of an action *would* be even if no such action takes place, which can be seen as a rudimentary form of imagination. In calling to mind the consequences of actions that never actually take place (albeit often unaccompanied by relevant phenomenology), organisms can represent *merely possible* scenarios. As organisms evolved more sophisticated ways of interacting with the environment, including actions directed at goals beyond the organism's immediate surroundings, the capacity for anticipating the results of possible actions also had to get more sophisticated, allowing for the *chaining together* of anticipated pairs of possible actions and anticipated sensory consequences into more and more elaborate action plans (Pezzulo & Castelfranchi, 2009; Pezzulo & Cisek, 2016).

Embodied accounts of imagination "constrain the space" of imaginative operations "to those that can effectively use forward models that were originally developed for online interaction" and, as a result, imagination "*retains essential features of* online interactions (i.e. forward models) although *it does not consist* in online interaction" (Pezzulo, 2017, p. 4, original emphases).[22]

5.3.4 From Action Imagination to the Environment

It's clear that merely anticipating the outcomes of one's own actions in the world can only get one so far. To successfully interact with a changing environment, one

[22] It is important to note that there is a different way in which imagination can be constrained by embodiment. Our imagination is sometimes constrained by *merely temporary aspects* of our present bodily state. A nice demonstration of this kind of constraint on the imagination is provided by Binet (1899, p. 29): open your mouth as wide as possible and keep it open, while doing so try to *imagine* saying the word 'bubbles'. Binet argued that doing so is impossible. This might be a bit too strong, but it certainly seems difficult, and significantly more difficult than doing so with one's mouth closed. Thus, in this and many other ways, the particular state that one's body is in when trying to engage in imagination can constrain what can be imagined. These temporary constraints on imagination differ from those that we are concerned with precisely because one can introspectively notice them by varying one's bodily state. However, the presence of these temporary constraints serve as a good example of the way in which bodily state can constrain imagination. The difference in the case of more general constraints that arise from the kind of creature one is lies in the fact that one cannot vary them (e.g., by temporarily becoming a different species) so as to notice the variation in imaginative capacity that results.

must also anticipate scenarios that are not the result of one's own actions. In order to explain the origins of imaginative capacities that go beyond our own ability to directly affect the world, it is important to turn to another important feature of motor control. Many of the kinds of action that we engage in involve interacting with dynamic rather than static features of the environment. In such cases, if one wants to effectively coordinate action, it is not sufficient to merely anticipate one's own effects on the environment, one must also anticipate the way the environment will change over time as one carries out the given action (see Burr & Jones, 2016; Pezzulo & Cisek, 2016).

It might be helpful to briefly (and tentatively) look at imagination from the perspective of the *predictive processing* framework (Clark, 2013, 2016).[23] Predictive processing is the view that perception, imagination as sensori-motor simulation, and psychological phenomena in general "come about through the same process: minimization of prediction error" (Kirchhoff, 2018, p. 754). Roughly, our brain makes a prediction of what the consequences of certain actions or interactions with the environment are and then corrects these predictions on the basis of the feedback from the actual consequences. What is important for us is that on these views, imagination is closely related not only to our own actions, but also to our *embedding environment* (see Clark, 2017; Kirchhoff, 2018; Jones & Wilkinson, 2020). As Kirchhoff puts it, "minimization of prediction error is not restricted to the brain alone but involves the entire organisms (morphology, action capacities, and so on) *and its embedding environment*" (2018, p. 761, emphasis added).

Relating this back to our discussion of the fact that imagination has to be restricted in order to be epistemically useful, we get the following overall picture. The "core idea, that imagination involves possible actions and experiences, generates constraints that come from two main sources. The first is bodily constraints that are the result of the organism's phenotype. The second is constraints from past experience" (Jones & Wilkinson, 2020, p. 105). Importantly, and in line with the main claim of this chapter, Jones and Wilkinson continue by pointing out that this means that imagination and the relevant constraints "are determined by the shape of one's body" (ibid.).

Even though the predictive processing framework helps to make this explicit, presumably on any moderate embodied account sensori-motor recreation will involve some representation of the kind of ways that the environment changes. So, imagination, on an embodied account, is able to go beyond the effects of our own actions through its being shaped by the embedding environment of the agent. Imagination reliably recreates the "environmental causes" because the environmental embedding is part and parcel of the embodied perspective on cognition in general (and

[23] See Kirchhoff (2018) and Jones & Wilkinson (2020) for excellent discussions of imagination in such frameworks. Jones and Wilkinson their work is particularly interesting as they discuss a version of the predictive processing framework that are closely related to the sensori-motor simulation view discussed here.

in the case of, for example, predictive processing, on perception and imagination in particular) (Kirchhoff, 2018, p. 756).

5.4 Embodiment, Introspection, and Objectivity

With this theory of embodied imagination on the table, let us return to the two issues that we discussed in relation to the appearance-based theories of Balcerak Jackson (2018) and Gregory (2019): the problem of introspective accessible grounds of the justificatory role of imagination and the problem of objective evidence. We will see that on an embodied theory of imagination, both these worries are alleviated.

Before we discuss these issues in relation to the embodied theory of imagination, remember the note above about the relation between the appearance-based theories and the embodied theories. As a theory of imagination, the latter seems to subsume the former in that both take imagination to be the recreation of perception, yet the embodied theories also stress the importance of *motor-cognition*. In particular, the embodied theory of imagination focuses on the interaction between perception and motor-cognition in terms of *sensori-motor simulation*. This allows us to explain not only the fact that we can imagine what it might look like to see an apple hanging from a tree, we can also imagine grasping that apple and the effect of us moving forward on the branch of the tree has on our grasping the apple (Pezzulo & Cisek, 2016). However, there is an additional feature of embodied approaches, namely adopting the particular perspective of the embodied cognitive sciences. For example, taking a particular view on embodied cognition (e.g., à la Barsalou), we can even extend this kind of recreative imagination to explain *propositional* imagination (e.g., imagining that there is a tiger), whereas the appearance-based approach only focuses on *objectual* imagination (e.g., imagining a tiger) (Yablo, 1993; Balcerak Jackson, 2018).[24]

So, the embodied theory of imagination, as a theory of imagination, suggests that we can recreate more cognitive features (in particular, aspects of motor cognition). In addition, there are 'kinds' of imagination whose epistemic usefulness embodied imagination theorists can explain that fall outside the scope of the appearance-based theories (e.g., propositional imaginings). This is important, for it suggests that theorists such as Balcerak Jackson and Gregory could also adopt some aspects of the embodied approach to imagination, without necessarily having to adopt all of them.

[24] That is, on certain theories of embodied cognition, propositional content also reuses the cognitive machinery we use for online cognitive activities such as perception, et cetera. Though the arguments of this chapter hold independently of this view.

5.4.1 Imaginative Blocks, Introspection, and Embodiment

Recall that an imaginative block is something that you (currently) fail to imagine. A problem arises when we fail to imagine something when there are unchosen (or architectural) constraints in place. In those cases, we might not be aware of the effects of the constraints due to the fact that they may fundamentally be introspectively inaccessible. We saw that this suggests that whatever it is that grounds the epistemic justificatory force of imagination should not, and cannot, be introspective seemings. This is not a result of the fallibility of introspection or because introspection is particularly unreliable for the case of imagination. It is because the constraints on the mechanisms that give rise to experiences might fundamentally be introspectively inaccessible.

On the embodied approach to imagination, imaginative blocks still arise and they still present the same epistemic problem: there might be no way for us to assess the effects of such constraints introspectively. Our imaginative capacities are partly constrained by the nature of our embodied systems, regardless of whether imagination involves conscious mental imagery, and thus we won't be able to learn about such constraints through merely reflecting on the nature of the conscious mental imagery that is only sometimes a feature of imagination. However, the point is that on an embodied theory of imagination this does not come as a surprise. In fact, it is in line with the findings of embodied cognitive science that our embodiment might have surprising effects on our cognitive lives; effects that we would not have expected based on our introspective reflections. That is, embodied imagination *explains* why the problem of imaginative blocks arises.

The crucial difference is that on the embodied imagination theory it is clear from the beginning that, in order for embodied cognition to provide us with a new way of understanding the constraints on imagination, it is necessary to move beyond traditional philosophical introspective analysis and turn to the scientific study of the mechanisms that support imaginative activity. What provides imagination with the justificatory force that it has are the constraints that imagination inherits from the cognitive machinery that it uses for recreation. We should not expect to be aware of these features of imagination (and their limits) based on reflecting on our introspective seemings (or, as Gregory put it, "to look and see"). We should turn to (embodied) cognitive science to tell us what the mechanisms of imagination are, what constraints are in play, and what the limits are of the justificatory role of imagination (based on the answers to the two former questions).

In a sense, Nichols' (2006a, p. 247) suggested solution to the (original) problem of imaginative blocks already hints at this. The following example nicely shows this:

> Chimpanzees can't make anything out of the proposition that the set of real numbers is finite. But obviously a cognitive ethologist can perfectly well make this observation about chimpanzee cognition without having to 'make something of the thought' that the set of reals might be finite.

The gist of Nichols' observation is that, even though the chimpanzees themselves are not aware of, nor (it seems) can introspectively access, their cognitive limitations, the ethologists studying primates can explain this. In our discussion, this comes down to the fact that we might not be able to introspectively access the features of our imagination, but we as modal epistemologists "are merely obligated to explain *how* the blocks arise" and how imagination works (Nichols, 2006a, p. 248, emphasis added). As modal epistemologists, we cannot appeal to our experiences as imaginers, but we can appeal to the embodied cognitive sciences to explain the characteristics of imagination that ground its justificatory role.

In more contentious terms, the perspective shift of an embodied approach to imagination suggests that the epistemology of recreative imagination will most likely be externalist.[25] By heavily relying on what the recreative aspect of imagination is (the simulation of sensori-motor activation), the embodied imagination theorist suggests that what grounds the justificatory role of imagination is the fact that we recreate sensori-motor activity that the agent would use if they were to actually perform the imagined actions. Clearly, this fact need not be internally accessible to the agent in question.

5.4.2 Sensori-motor Simulations and Modal Objectivity

The problem of modal objectivity, for the appearance-based theorists, concerned their focus on *phenomenal* evidence. Their focus on evidence in the form of *possible experiences* meant that they had to make strong assumptions about experiences of possible subjects in possible worlds, in order to suggest that we can use imagination to gain justification for beliefs in what is objectively possible. Preferably, we would have an account of imagination that does not need to rely on such an assumption while securing the justification for beliefs in objective modal facts. Embodied imagination can do just this.

The reason why embodied imagination is more straightforwardly related to objective modal facts than the appearance-based theories (as they are currently phrased) can be highlighted in two, related, ways.

First of all, the interpretation of the cognitive mechanisms that imagination recreates are geared towards more objective features of reality from the start (this is the perspective shift that comes from embodied cognition). For example, perception, in embodied cognitive science, is not viewed as a static process concerning static perceptual experiences. Instead, perception is thought of as an active process, relating closely to potential actions the environment allows us (e.g., Gibson, 1979; Noë, 2002; Grush, 2004). Instead of thinking that 'online' perception gives

[25] Though, as mentioned before, BonJour (2003, sec. 5) gives a defense of internalism in general based on similar 'seemings'. Moreover, there are internalist options that are less reliant on introspective access: e.g., Wright's (2004; 2014) theory of rational entitlement.

us phenomenal experiences, embodied cognitive science already takes perception to be related to objective facts about reality, for example in terms of *affordances* (see also Nanay, 2011a,b; and Strohminger, 2015). Without going into too much detail, such accounts suggest that perception "does not begin with a static retinal array, but with an organism actively moving through a visually rich environment" (Wilson & Foglia, 2017, §2.4). This, in turn, has as a result that perception is "used to distinguish *agent-dependent* and *objective* features of one's environment" (ibid., emphases added). Strohminger (2015) even suggests an epistemology of *objective* modality based solely on this feature of 'online' perception. In taking imagination to be the recreation of perception, the embodied view suggests that instead of providing us with possible experiences, imagination already provides us with possible agent-dependent and objective features of the environment. That is, embodied imagination gives us possible actions that the simulated environment allows us; it provides us with evidence for possible affordances. The other major cognitive capacity that imagination recreates on an embodied account is that of motor action itself. Such motor simulations straightforwardly relate to the possible, objective actions and possible effects thereof on the environment (see Grush, 2004; Jeannerod, 2006; Pezzulo & Cisek, 2016).

However, not everyone might accept that perception is related to the objective world through, e.g., affordances. Still, even if one rejects the above interpretation of perception, one can argue that embodied imagination latches on to objective possibilities as opposed to merely possible sensations. The argument involves considering the evolutionary purpose of imagination on such accounts. The evolutionary purpose of imagination is that we use it to 'detect' possible opportunities and risks that allow us to change our actions accordingly. In order to be a useful tool for this, it makes no sense that imagination operates completely independently of our knowledge of what we take the world to be like. That is, sensori-motor simulations are use, not for what it would be like to perform certain actions, but for *what the world would be like* as a result of them (Nichols, 2006a; Pezzulo & Castelfranchi, 2009; Pezzulo, 2011; Kroedel, 2012; Langland-Hassan, 2016; Pezzulo & Cisek, 2016; Williamson, 2016a). Pezzulo explicitly discusses this evolutionary side of imagination: "living organisms are selected by evolution to produce good indications for action, since accuracy of their internal modelling processes is necessary for the success of their schemas, and ultimately for their survival" (2011, p. 91). Embodied imagination gets us this kind of knowledge by allowing us to test the effects of our actions on our environment without actually engaging in the relevant actions. These are real, objective possibilities that are tested, not merely sensations thereof. In general, imagination allows us to "get an epistemic grasp over the external reality, [...] *in terms of action possibilities and action goals*" (Pezzulo, 2011, p. 87, emphasis added).

Imagination from the perspective of embodied cognitive science (i) explains the issue of imaginative blocks and push towards an appeal to embodied cognitive science to explain the justificatory role of imagination and (ii) provides us with evidence of *objective* possibilities, rather than phenomenal evidence of possible experiences.

Let me stress again that appearance-based theorists can use similar solutions to the problems of imaginative blocks and modal objectivity *without* adopting a full-blown embodied account of imagination. For example, instead of appealing to introspective seemings, they might appeal to the cognitive science concerning perception in order to explain the justificatory force that imagination has. Additionally, if they adopt an account of perception that is more closely related to the objective features of our environment (as those discussed above), their account would also provide evidence of objective possibilities without the controversial assumptions that they currently need to appeal to. That is, they could appeal to the scientific explanations of the neuro-physiological works of our perceptual system in order to ground the justification of imagination for objective possibilities.

5.5 Limits of Embodied Imagination-based Epistemologies of Possibility

Looking at recreative imagination from an embodied cognitive science perspective allows us to overcome the problem of modal objectivity because imagination is in the business of providing us justification for possible actions and the effects thereof given a particular environment. However, as I will discuss in this concluding section, the source of this virtue is also responsible for embodied imagination's biggest limitation. As we will see, the situations we can justifiably believe to be possible based on embodied imagination are rather limited. Yet, before we turn to this discussion, let me briefly mention one additional virtue of the embodied imagination approach.[26]

Remember that in this dissertation we are looking for a cognitively plausible epistemology of possibility while accepting methodological naturalism (see Williamson, 2007; Nolan, 2017). That is, the (philosophical) methods that make our beliefs in what is possible justified are roughly "of the same general kind and [are] generally harmonious with the methods of the sciences" (Nolan, 2017, p. 8). When Nolan discusses imagination in particular, he suggests that turning to the literature on *perception and affordances* will be a promising starting point for a respectable imagination-based epistemology of modality in line with methodological naturalism (see also Phillips et al., 2019).[27] Embodied imagination seems a prime candidate for such a methodologically naturalistic imagination-based epistemology of possibility.

[26] This virtue applies equally well to the appearance-based approaches.

[27] "The second area of the psychology of modal judgements is one that I am currently less familiar with, but which has attracted the attention of a number of philosophers of mind: the study of perception of *affordances*. [...] This perception of opportunities and options and possibilities, and non-perceptual beliefs about these features of our surroundings, seem to be a relatively basic part of our epistemic repertoire, and seems to be providing modal information, or at least dispositional information" (Nolan, 2017, p. 20, original emphasis).

It appeals to the sciences to inform us about the features and limits of our imaginative capacities and it relies on cognitive capacities crucial for our everyday life (e.g., perception and motor control).

5.5.1 Embodied Nature and Modal Modesty

Accepting embodied imagination means that we have to accept that we are stuck imagining from our own embodied perspectives. This entails, in turn, that some forms of knowledge will always be beyond justification using imagination. Our perceptual and motor abilities evolved to cope with certain kinds of environments and the kinds of actions that organisms like ourselves tend to engage in such environments. So if our imagination is shaped by similar constraints as a result of reusing the same systems, we should expect reliability when we use our imagination to address situations that are close to our everyday interactions with the world. But, as Nichols pointed out, when "the psychological systems are being used outside their natural domain [. . .] there's less reason to think that they will be successful guides in [such] foreign terrain" (2006a, p. 253).

We should focus on using imagination within its natural domain; the natural domain being restricted by our embodiment. What this means is that we can only expect imagination to be a reliable source of knowledge in domains such that our embodied constraints were shaped to deal with them (e.g. evolutionarily or developmentally familiar settings). In terms of the epistemology of possibility, the scope of the kinds of modal claims that we can come to know through imagination will be limited. The instances where imagination can be a reliable guide to possibilities are those mundane possibilities ('close by') that are similar to situations that we encounter in our ordinary life. However, if we go out of this familiar domain and consider more exotic or 'far away' possibilities, our imagination becomes less reliable (or perhaps collapses into only providing phenomenal evidence). So, although sensori-motor simulation gets us knowledge of objectively possible actions and the effects thereof, it is only of our environments that we can gain modal knowledge. The positive story of getting *objective* modal knowledge and the resulting modal knowledge being naturally limited go hand in hand.

The scope of the reliability of imagination is likely to be vague and without a sharp cut-off point. For example, when looking at my table and seeing a book lying on it, I can imagine picking up the book, throwing the book, et cetera. My sensori-motor simulation provides me with evidence of the corresponding objective possibilities (e.g., it is possible that I pick up the book, etc.). These all concern possibilities that *I* can do. Perhaps, I can even project these kinds of imaginings into what *you* could do. For example, I see you standing near a table with a book and through sensori-motor simulation I come to believe that it is possible that you throw the book. But this raises the question: how far does such a projection go? Can I imagine myself to be a little bit taller than I actually am? Can I imagine giving birth? The former, intuitively, seems possible, but the latter arguably not (Balcerak Jackson, 2016, p. 47). Similarly, can I reliably project sensori-motor abilities to those with significantly less or more skill (Pezzulo et al., 2010)? What the boundaries of our embodied experience and imagination are is something that lies beyond the scope of this chapter. I merely want to point out that it is very likely that these boundaries are on a continuous scale. For example, a congenitally blind person might have developed the ability to use vocal clicks to navigate their environment. If they were to imagine that they locate a particular object in a completely blacked-out room through vocal clicks, they might, based on this imagining, reliably conclude that it is possible. However, given that I am not able to use vocal clicks, it seems that if I were to imagine that I use these to locate an object in a blacked-out room, this imagining would *not* be a reliable guide to what is possible.

Thus, on the embodied approach we should expect imagination to yield knowledge of modality only in *restricted* circumstances. For example, it seems that we should expect our imagination to only yield reliable modal knowledge about *physical possibility*, since our sensori-motor systems evolved to guide actions that are governed by the laws of physics in this world, not some exotic alternative laws of physics in a distant possible world.[28] Yet, this may be a best-case scenario. The lesson to be drawn from the embodied cognition literature is that the kinds of bodies that we possess will constrain our imagination in *unexpected* ways. Many physical laws will be irrelevant to the way that we, as humans, interact with the world, and as such our imagination may only be reliable in relation to the particular physical possibilities that happen to correspond to the laws that are important to *our* forms of interaction. For example, at the scale that we ordinarily engage with the world, the laws of quantum mechanics have little impact, so we should not expect our imagination to provide reliable access to knowledge of quantum possibilities. That is, an embodied imagination-based epistemology of possibility is *perspectival* in the sense that it is reliable when it comes to possibilities that *concern us*.[29]

On the theory of embodied imagination as developed in this chapter, the resulting knowledge we get (i) is properly grounded in the findings of (embodied) cognitive science as opposed to introspective seemings; (ii) relates to *objective* modality as opposed to phenomenal evidence; and (iii) is *modally modest* in the sense of Van Inwagen (1998) and Hawke (2011), the kind of modal knowledge that we can get justification for clearly delineated.[30] A crucial feature of embodied imagination is

[28] Remember the discussion of the different kinds of modality in Chapter 1 (Section 1.1).

[29] That is, based on the sensori-motor account that we use (e.g., of Pezzulo and colleagues), the cases where we can get modal knowledge are the cases that either concern *movements* and *abilities* of our bodies or effects of such movements and abilities within the environment we've developed. It thus seems that embodied imagination is ideally suited to provide an epistemology of so-called agentive modalities (see, e.g., Maier, 2015).

[30] See Chapter 11 for more on modal modesty and the problem of delineating the modesty without collapse into radical modal scepticism.

that precisely *how* our imagination is constrained and what impacts this has on the epistemology of imagination is unclear until more empirical work has been done.[31]

It would be interesting to see how far the epistemology of modality can be pushed based on an embodied-imagination account as provided here. In particularly, in response to the *integration challenge* (i.e., aligning one's metaphysics of modality with one's epistemology thereof), the embodied imagination epistemology of modality seems very well suited to be linked to *potentiality-based* accounts of modality (Vetter, 2015). Relatedly, I think that a theory of embodied imagination with a focus on potential actions and the effects thereof, as the one proposed here, might play an interesting role in an epistemology of causation in combination with proprioceptive experience of causal forces (Anscombe, 1975). Both of these questions are left for future research.

[31] Thanks to Max Jones for reminding me to stress this point.

Part II

Similarity

Chapter 6

Introduction to Similarity-based Theories

In the first half of this dissertation, we discussed what, according to many, is the main (or 'traditional') method of gaining knowledge of possibilities used by modal empiricists: imagination. However, through the different chapters, we saw that it is not obvious that the different ways in which one can cash out what they mean by 'imagination' are proper bases for an epistemology of possibility. In Chapter 5, we concluded that the most promising imagination-based approach gives rise to knowledge of a limited number of possibility claims. In the second part of the dissertation, I will focus on *similarity-based* approaches to the epistemology of possibility (Roca-Royes, 2007, 2017; Hawke, 2011, 2017), which I think have been under-appreciated.

To give a sense of how under-appreciated this approach has been, we can turn to discussions in the main overview articles on the epistemology of modality.[1] The Stanford Encyclopedia of Philosophy entry on the epistemology of modality (Vaidya, 2016) only briefly discusses similarity-based approaches and only mentions the work of Roca-Royes (2017). Strohminger & Yli-Vakkuri (2017) focus mainly on imagination-based, counterfactual-based, and rationalist deduction-based approaches to the epistemology of modality and only mention Roca-Royes in a short phrase as 'one of the other methods'. The same goes for Vaidya & Wallner (2018).[2] Similarly, when Vaidya (2017) discusses empiricist approaches to the epistemology of modality there is no mention of similarity-based approaches. Finally, in Mallozzi's (2019) introduction to a special issue of *Synthese* on new directions in the episte-

[1] Overview articles from before 2011 (McLeod, 2005; Evnine, 2008) do not mention these approaches at all. However, since Roca-Royes (2007) and Hawke (2011) are the first ones, as far as I know, that explicitly discussed a similarity-based approach and the most 'popular' discussion of a similarity-based approach is that of Roca-Royes (2017), this is to be expected.

[2] Though, to be fair, Vaidya and Wallner focus on conceivability-based, counterfactual-based, and rationalist deduction-based approaches only "as a function of [their] goal" (2018, p. 3).

mology of modality, she only discusses imagination-based and counterfactual-based approaches within modal empiricism and, similar to Strohminger and Yli-Vakkuri, only mentions Roca-Royes' work in passing.

Very roughly, similarity theorists hold that our knowledge of actuality provides us with justification for our beliefs about what is possible if the objects or events involved are *relevantly similar*. For example, I have a wine glass of which I believe that it *could* break; that is, I believe that it is possible that the wine glass breaks. According to the similarity theorists I am justified in having that belief due to the fact that I believe that this wine glass is (relevantly) similar to another wine glass I once had that *did* break.[3] Or, to use an example from Roca-Royes:

> I know that the wooden table in my office, Messy, is not broken. How do I know that? I see it. Although not broken, Messy can break. How do I know that? Because the table I had before Messy, which we may call 'Twin-Messy', was a twin-sister of Messy, and it broke; and I know that Twin-Messy broke because I saw it. (2017, p. 226)

This view is intuitively plausible and promises to ground knowledge of non-actual possibilities in our knowledge of actuality. Hawke (2011) and Roca-Royes (2017) are the two *loci classici* of similarity-based approaches to the epistemology of possibility,[4] but recently some more work has appeared (Hawke, 2017; Leon, 2017; Dohrn, 2019). In this introduction, I will discuss, in some detail, the theories of Hawke and Roca-Royes and one of the main objections against similarity-based approaches: they fail to specify exactly what their central notion, *relevant similarity*, is (Hartl, 2016; Vaidya, 2016). This objection will set the stage for the two chapters in this part of the dissertation.

6.1 Hawke's Safe Explanation Theory

Hawke (2011, 2017) presents, what he calls, a *safe explanation theory* of the epistemology of possibility. His theory is heavily empiricist and is motivated by a defence of *modal modesty*, such as that of Van Inwagen (1998). What Hawke tries to do, in a sense, is to explicate Yablo's (1993) conceivability-based approach in such a way that Van Inwagen's (1998) argument is no longer susceptible to the objections of Geirsson (2005). Geirsson argues that one of the arguments that Van Inwagen gives in favour of modal modesty relies on an interpretation of Yablo's work that

[3] This relies in part on the actuality principle. Recall that we discussed this in Chapter 1 (Section 1.4.2): whatever is actually the case is possible.

[4] Let me stress that both Hawke and Roca-Royes have explicitly resisted committing themselves to the claim that something like the similarity principle is the *sole* ground for possibility knowledge (Hawke, 2017; Roca-Royes, 2019b).

is too demanding, resulting in a theory of conceivability that would be impractical – i.e., that would predict that our knowledge of ordinary possibility statements would not be justified because of it. Hawke (2011) presents an interpretation of Yablo-conceivability such that we both are justified in believing ordinary possibility statements on the basis of it, but we can keep Van Inwagen's argument for modal modesty.

Hawke suggests that in order for a conceived situation, s, that represents a proposition, p, to provide justification for the agent to believe that p is possible, the elements of the conceived situation should be less "modally controversial than" p (2011, p. 359). That is, there should be independent justification that the elements in the conceived situation are possible. One way this might be is if the proposition of interest, p, is a logical implication of some other propositions, $q_1, \ldots, q_n$, and these other propositions are all *modally less controversial* than p. We then say that these propositions form a *modally safe explanation* of p (see Hawke, 2011, p. 359). However, a worry arises of an infinite regress of more and more modally less controversial propositions.[5]

> There is a nagging worry about the 'safe explanation' theory: the account calls for justification of possibility-claims in terms of other, already justified possibility-claims. As it stands, this could either lead one in a circle or upon a path of endless justification. (Hawke, 2011, p. 359)

Roughly, Hawke holds that there is a *base set* of propositions that are *basic* possibility claims. That is, there is a set of propositions such that it is reasonable to believe without question that they are possible (see the discussion of modal Mooreanism in Chapter 1, Section 1.2). Let's call this set $B = \{q_1, q_2, \ldots, q_n\}$. A proposition is in the base set only if (i) it is true at the actual world or (ii) if it is *relevantly similar* to an actual r.[6] The totality of our modal knowledge is *recursively* defined on this base set and includes all the propositions that are 'safely explained' by the propositions that are in this base set. We are justified in believing the possibility of any proposition that *logically follows* from the base set.

What is crucial for our purposes is to understand how Hawke interprets 'relevant similarity'. Let us first look at what Hawke calls the *similarity principle* and then look at what he takes to be involved in this principle. The similarity principle, according to Hawke, is the following:

[5] Hawke notes another worry, namely that of combination: there is no guarantee that the conjunction of two possible propositions is itself possible (see also Hawke, 2017, p. 296).

[6] In later work, Hawke (2017) suggests some further propositions that may be in the base set. A proposition may also be in the base set if (iii) it is part of the best explanation of an established fact, or (iv) it is a combination of two independently existing, reasonably believed to be possible, states of affairs.

Hawke's Similarity Principle (HSP): "If two things (situations, objects) are similar in some respects, then the possibilities (relevant to the similarities) concerning those things are likely to be the same." (2011, p. 360)

The problem of similarity, in general, is that (almost) anything is similar to anything else "in some respect." Therefore, similarity theorists talk of *relevant* similarity. Hawke provides some information as to *which* properties he thinks the similarity principle should be applicable to. He puts this as follows:

> What counts as 'relevant' similarity when it comes to making judgements of possibility? It would seem that a similarity is relevant to the possibility of p if that similarity stands in some kind of *causal or determining relation* to the advent of the states of affairs that make p true. [. . .] Indeed, much more needs to be said to properly explore and evaluate the similarity principle. (2011, p. 361, emphasis added)

As Hawke himself admits, "much more needs to be said" here and this is indeed one of the main arguments raised *against* similarity theories in general (Vaidya, 2016) as well as Hawke's theory in particular (Hartl, 2016). I will discuss Hartl's objection in the last section of this chapter. Let me finish the discussion of Hawke's theory with a brief remark on his justification for accepting (**HSP**). Hawke presents an inductive argument for the use of the similarity principle. He notes that the actuality principle is the ultimate test of whether or not something is possible: when the similarity principle predicts something to be possible, we can test it with the actuality principle by (trying to) actualise the possibility.[7] According to Hawke, examples where the similarity principle is tested by the actuality principle

> are, clearly, innumerable and the similarity principle, I am sure it is agreed, tends to fair [sic] very well in the face of such tests. What this amounts to is that significant evidence exists for the truth of the similarity principle. Thus, one may conclude inductively that the similarity principle is true. (2011, p. 361)

Such an inductive justification for the similarity principle is, as we will see in the next section, also appealed to by Roca-Royes (2017) – the other main champion of a similarity-based epistemology of possibility.

[7] Note that the claim is not that we can test *all* predictions made by the similarity principle, sometimes trying to actualise the predicted possibility might be practically or technologically very hard (if not currently impossible). For example, based on the similarity principle, I might predict that it is possible for humans to colonise Mars, however, this is not a possibility that we can (easily) actualise. Thanks to Peter Hawke for encouraging me to clarify this and for the example.

6.2 Roca-Royes' Similarity Theory

Whereas Hawke had, somewhat, theoretical motivations for developing his similarity theory, Roca-Royes (2017) starts from pre-theoretic intuitions about how we acquire knowledge of ordinary possibility statements.[8] Remember, for example, the table example mentioned above. Taking this pre-theoretic description, she believes "that, roughly, *this is how* we form informed judgements about unrealized possibilities that are both accessible and basic. [...] I believe, more importantly, that such route to modal judgement is knowledge-conferring" (2017, p. 226, original emphasis). So, according to Roca-Royes we rely, just as with Hawke's theory, on the actuality principle (we know that Twin-Messy could break, because it did break) and on a similarity principle.

Two objects are relevantly similar, according to Roca-Royes, if they are *epistemic counterparts*. This is because "any two entities that stand in the counterpart relation do so in virtue of being similar in some relevant respect" (2017, p. 226). If we know a realised (actualised) possibility of one of these objects, then we can extrapolate this to the relevantly similar object – i.e., to its epistemic counterpart – regardless of whether we know that the epistemic counterpart actually has the property in question or not.

Of course, Roca-Royes is aware that more needs to be said here and she does so by spelling out two important instances of *prior knowledge* that are involved in this kind of reasoning: (i) we have prior categorical knowledge and (ii) we have some prior nomic knowledge.[9] These two pieces of prior knowledge together give rise to our similarity judgements. The nomic knowledge that we require are law-like principles of the following form (Roca-Royes, 2017, p. 230):

$$P(x) \to \Diamond Q(x)$$

What this principle (schema) is supposed to capture is "the idea that causal *powers* and effect *susceptibility* depend on qualitative character" (idem, p. 229, original emphases). Instances of this principle are examples such as 'tables can break' and, more explicitly in conditional form, "if an entity has a heart, it *can* die of a heart attack" (idem, p. 230, original emphasis).

What Roca-Royes calls the required categorical knowledge is knowledge of the antecedent of the nomic principle. For example, that an object is a table or that an object has a heart. Roughly, in terms of Chapter 8, this requires us to *categorise* objects: we need to judge that object a is a human and that object b is so too.

[8] Roca-Royes has a very elaborate account of our modal knowledge that is *non-uniform*, in the sense that we might use different methods to justify different kinds of modal knowledge. She also has provided a number of convincing arguments for this position. See Roca-Royes (2010, 2017, 2019b, forthcoming).

[9] She also mentions that we have prior knowledge that actuality implies possibility. Again, we ignore this to focus on the similarity-aspect of her theory.

In general, the prior categorical knowledge and the prior nomic knowledge together give rise to the similarity-based epistemology of possibility.[10] How detailed the categorical knowledge needs to be in order for the nomic principle to result in reliable predictions is something that needs to be spelled out. Very roughly, Roca-Royes suggests that when we rely on *few* objects as our source, then the categorical knowledge should be as detailed as possible. For example, if we consider whether a table can break based on *one* other table that broke, then it would be best to know that it was a particular sort of IKEA table. On the other hand, the more tokens (and especially the more *varied* tokens) we have as our source, the more general the categorical knowledge can be. If we know of many different tables that they broke, it will be enough that the next object is also a table (independent of what kind of table it is). In these cases the reasoning will be more like (enumerative) induction.[11]

Roca-Royes provides three ways one might, roughly, justify the use of such ampliative methods (as the categorical and nomic knowledge her account relies on). First she notes that we use ampliative methods all the time, denying the use here would result in being forced to give up ampliative methods in general, resulting in widespread scepticism (a roughly Williamsonian 2007 defence). Secondly, she points out a possible justification of such ampliative methods, namely the appeal to *entitlement of rational deliberation* to justify the epistemic foundations of these ampliative methods (we will discuss this in more detail in Chapter 8, Section 8.3.3). Finally, she notes that we could test these methods' predictions by trying to actualise them, each time this works, it would be inductive evidence for the reliability of such similarity-based judgements (this is similar to Hawke, 2011).

Based on the nomic principle, the kind of similarity that is needed between objects concerns properties that could be used in an antecedent of a nomic principle. As Roca-Royes puts is, "the similarity at issue is similarity in categorical *intrinsic* character" (Roca-Royes, 2017, p. 233, emphasis added). In particular, she focuses on the notion of *qualitative anchor*, which describes "those [properties] (appearing in true grounding principles) capable of playing the epistemic role of allowing us to (groundedly) transition to a given *de re* possibility" (idem, p. 237).[12] Note that

[10] Roca-Royes' account is perhaps more accurately described as an *induction-based* epistemology of possibility (as she, as well as some commentators, have done). However, many seem to gloss over the fact that a crucial aspect of induction is a similarity judgement (perhaps due to a focus on the uniformity of nature problem).

[11] What is interesting to note here is that Roca-Royes goes on to suggest that this observation might track what Van Inwagen (1998) has in mind with his distinction between mundane and exotic possibilities: the former are instances where we have many and varied priors and the latter where we have few. She says that this is why we might be better at getting to know mundane possibilities – i.e., those of which we have experienced many prior (similar) instances.

[12] Additionally, she points out that we should focus on the combinations of P's and Q's where, with respect to $P(x) \to \Diamond Q(x)$, $P(x)$ is *temporally* prior to (the beginning of) $\Diamond Q(x)$. Something that she does not note, but which might also be the case in light of what will follow is that P might be more general than Q. For example, in virtue of being a house, it is possible to be a small

it is important that this is an *epistemic* notion, so one might be unaware of some features of the intrinsic character and such unawareness *does* affect the similarity reasoning.

According to Roca-Royes, the focus on the intrinsic nature of these qualitative anchors that are supposed to distinguish *relevant* similarity from 'random' similarity allows her to distinguish defective from satisfactory instances of similarity reasoning. The focus on properties that ground the inference to the possibility claim in the consequent of the nomic principle is, according to Roca-Royes, the thing "that explains the defectiveness of the reasoning in the second [bad] pair while leaving the first two [instances of good reasoning] in good standing" (2017, p. 236).

6.3 Relevant Similarity

From our discussion of the theories of Hawke (2011) and Roca-Royes (2017), we can distil a general description of the crucial similarity reasoning: we know some object, x, has a particular property, P. From this, we deduce that that same object, x, has yet another property, $\Diamond P$ (by the actuality principle: whatever is actual, is possible).[13] Then, we extrapolate that another, relevantly similar, object, y, also has that property, $\Diamond P$. This is the crucial *Similarity Argument*:

Similarity Argument (SA):

P1. x has property P.

C1. x has property $\Diamond P$. (actuality principle)

P2. x and y are relevantly similar relative to property P.[14]

C2. y has property $\Diamond P$. (from C1 and P2)

The crucial premise here is premise 2: *relevant* similarity relative to the property of interest. Let's call this the *similarity judgement*. This will be the focal point for this part of the dissertation. A well-known problem for theories of similarity (of any kind, e.g., in counterfactual conditional semantics; scientific representations; analogy; etc.) is that "[a]ny two things share infinitely many properties, and fail to share infinitely many others" (Lewis, 1983, p. 346). The challenge this raises for theorists relying on (SA) is that they need to develop a notion of *relevant* similarity that distinguishes between good and bad instances of the similarity argument. For example, my cat and my pillow are both black, both are soft to the touch, both are

house. So, if something else is a house, it could be a small house.

[13] I will use '$\Diamond P$' as sloppy notation for '$\lambda z.\Diamond P(z)$': an object being such that it could possibly have property P.

[14] Strictly speaking, the relevance should be to property $\Diamond P$, however, as will become clear in the next two chapters, we can skip the actuality inference and focus directly on similarity with respect to embedded property.

composed of atoms, et cetera. As my pillow is an artefact, I conclude that my cat could also be an artefact. This is clearly not good similarity reasoning. However, if I conclude based on the fact that my pillow and a t-shirt I see in the store are both black, made from cotton, come from the same store, etc., that it is possible for this t-shirt to be an artefact, then I seem to have engaged in good similarity reasoning. The challenge is to give an account of 'relevance' that captures the difference between former and the latter kind of similarity reasoning (this challenge will be discussed elaborately in Chapter 7). As Aronson points out, "[w]ithout constraints on what is to count as a relevant feature for matching," similarity reasoning fails to capture anything interesting, as "any two things could be said to be similar or dissimilar to any degree" (Aronson et al., 1995, p. 21; see also Goodman, 1972; Lewis, 1986; and Morreau, 2010).

So, making explicit what they take to be *relevant* similarity is the most pressing issue for similarity theorists. Current critics of similarity theories argue that Hawke and Roca-Royes have failed to do this so far (e.g., Hartl, 2016; Vaidya, 2016). For example, in a recent paper, Hartl elaborately discusses and criticises Hawke's similarity principle.[15] He rightly points out that "[t]o be able to apply this principle, it is essential to determine which properties of objects or events are relevant because, in some sense, virtually everything is similar to everything else" (2016, p. 286). As we have seen, Hawke suggests that the relevant properties are causal properties. Hartl objects to this suggestion. He notes that

> if we accept Hawke's suggestion and assume that relevant similarity in modal cases is only *causal similarity*, the Similarity Principle would have a narrow scope. If it worked, it could only reveal *physical* possibilities: propositions that are true in those possible worlds that have the same, or very similar, ontology and laws of nature as the actual world. However, this restriction of analogical modal reasoning seems to be arbitrary.
> (2016, p. 287, original emphases)

It seems to me that Hartl raises two, potentially independent, objections here: (i) the resulting epistemology of possibility would be too narrow in scope and (ii) this definition of what properties are relevant is arbitrary. Let me briefly say something about the first objection, before turning to the main issue: is Hawke's proposal of what relevant similarity is arbitrary? Considering the first worry, it is unclear that Hartl presents us with convincing evidence that the scope of the resulting epistemology of modality would be *too* narrow. First of all, note that this question is rather unfair to Hawke who, like Van Inwagen (1998), accepts a form of modal modesty (moreover, Hawke is very explicit that defending modal modesty is his motivation). Arguments to the effect that Hawke's view predicts us as not having

[15] Though Hartl does not discuss Roca-Royes' theory, to a large extent the spirit of the objection – what is the notion of 'relevant similarity' supposed to be – carries over.

knowledge of certain, exotic, possibility claims thus seem to be non-starters. This is exactly what the theory is supposed to predict. Relatedly, the objection begs the question of a uniform epistemology of modality, but it is not clear whether we should aim for such a uniform epistemology of modality (Roca-Royes, 2017, 2019b), nor is it clear whether Hawke himself thinks that we should. For note that if one supposes a *non-uniform* epistemology of modality, then the method of gaining justification for a particular aspect of our modal knowledge will indeed be limited to that aspect.

Finally, one might think that, independent of the preceding remarks, spelling things out in terms of causal similarity gets the scope of our modal beliefs exactly right. As Nichols (2006a), and others, have noted, our cognitive capacities are not evolved to deal with *metaphysical modality per se* and it seems much more plausible that we have evolved to deal with our immediate surroundings and the possible variations thereof (see Williamson, 2007; Pezzulo & Cisek, 2016; Kroedel, 2017; Phillips & Knobe, 2018; Phillips et al., 2019). Of course, as discussed in Chapter 1 (Section 1.1), practical, nearby nomic possibilities *are* also metaphysical possibilities. So, relying on our ordinary capacities seems fine to get knowledge of *these* metaphysical possibilities (as we will also see in Chapter 8). The worry, rather, is that our ordinary cognitive capacities are not evolved to deal with *purely* metaphysical possibilities, those that do *not* concern nearby nomic possibilities. Or, to phrase it slightly differently, the worry is that our ordinary cognitive capacities are not evolved to gain knowledge of *the full range* of metaphysical possibilities. Interpreting Hawke's proposal as cognitively plausible, it no longer seems to be the case that the scope is too narrow, but rather that it is just right.

The second worry derived from Hartl's objection is related to, as we saw, the main worry for similarity theories: what notion of relevance should we rely on in similarity-based epistemologies of possibility? According to Vaidya, this is the main question for such theories: "What specific details of relevant similarity does one need to know to be in a position to make the relevant inference?" (2016, §4.2). Hawke and Roca-Royes only make some preliminary remarks; suggesting that the relevance comes from the causal relations or categorical intrinsic properties, respectively. Hartl argued that, at least the suggestion of Hawke, was arbitrary. But why so? What else should determine relevance?

In the first chapter of this part, Chapter 7, I will explore the literature on analogical and similarity reasoning (e.g., Hesse, 1966; Gentner, 1983; Vosniadou & Ortony, 1989b; Bartha, 2010) in order to determine how to distinguish between similarity simpliciter and *relevant* similarity. I evaluate different proposals as a basis for a similarity-based epistemology of possibility and it turns out that Hawke's suggestion to focus on causal relations is indeed one of the most plausible ways to cash out relevance for successful similarity reasoning.[16] In Chapter 8, I will develop a

[16] Whether or not this is what Hawke had in mind is an open question, but at least it seems that we justify his suggestion in hindsight by appeal to this literature

positive, similarity-based epistemology of possibility. In a sense, the theory is very closely related to Roca-Royes' (2017) in that it involves reasoning steps that are similar to her categorical and nomic knowledge. However, I will specify very precisely what I take relevant similarity to be, which will both be cognitively plausible and knowledge-conferring.

Chapter 7

Relevant Similarity, Predictive Analogy, and Causal Knowledge

Distinguishing different kinds of similarity is essential to understanding learning by analogy and similarity

– Gentner, 1989

In this chapter, we look at the research done in the field of analogical reasoning, where we find a broad spectrum of many different kinds of similarity relations that one can use in similarity reasoning (see Gentner, 1983; Gentner & Markman, 1997). A crucial feature of an account of similarity reasoning is its ability to distinguish between good and bad similarity arguments. One promising way of doing so is by relying on the predictive analogy similarity relation (Bartha, 2010). This similarity relation takes *relevant* similarity to be based on shared properties that have causal relations to the property of interest. I argue that if we base our epistemology of possibility on similarity reasoning reliant on the predictive analogy similarity relation, we require prior knowledge of the specifics of these causal relations. This is potentially problematic for similarity theorists; how so depends on one's account of causation. I suggest that properly developing the notion of 'relevant similarity' leads similarity theorists to a significant crossroads for their epistemology of possibility: either (i) predictive analogies are used in their epistemology of possibility and similarity theorists have to accept that the significant work is delegated to the epistemology of causation, with all the consequences of the particular theory of causation one accepts, or (ii) they need to develop an alternative to predictive analogy as a plausible ground for similarity reasoning.

7.1 Similarity Theories and Similarities

Remember that similarity-based epistemologies of possibility rely on similarity arguments, discussed on page 123. Such similarity arguments are a particular instance of a more general argument:

General Similarity Argument (GSA):

P1. x has property P.

P2. x and y are relevantly similar relative to property P.

C2. y has property P. (from P1 and P2)

We know that a particular object a has a particular property, P, and we extrapolate that another, relevantly similar, object b also has this property. Clearly, a lot hinges on how one cashes out the notion of 'relevant similarity' and some have pointed out that the current similarity-based epistemologies of possibility fail to properly specify what they mean by it or fail to argue for the kind of relevance that they rely on (Hartl, 2016). In this chapter, I turn to the vast research that has been done on similarity-based and analogical reasoning (e.g., Hesse, 1966; Gentner, 1983; Helman, 1988; Vosniadou & Ortony, 1989a; Falkenhainer et al., 1990; Chalmers et al., 1992; Bartha, 2010, 2019), something that the literature on similarity-based epistemologies of possibility currently lacks. I will provide a synthesised overview of the discussions in this field in order to search for an appropriate notion of *relevant* similarity for similarity-based epistemologists of possibility such that the reasoning from (SA) is plausibly cogent.

As I will be talking about 'similarity' in many different contexts, let me make a number of terminological distinctions to keep things clear. First of all, I will use *similarity reasoning* to talk about reasoning that is based on the (general) similarity argument discussed above. Secondly, I will use *similarity judgement* for the judgement of **P2** in (GSA). Thirdly, I will use *similarity relation* to talk about the particular relationship between x and y that grounds making the similarity judgement. Finally, I will sometimes use *similarity theorists* to talk about proponents of similarity-based epistemologies of possibility.

7.1.1 Domains and Analogies

Similarity reasoning is an ampliative method intended to extend knowledge. I will use the phrase 'domain', as is suitably abstract to include concrete objects, situations, hypotheses, complex systems, etc., for the source and target of such ampliative reasoning. In our examples so far, we have focused on similarity reasoning concerning single objects: object a is relevantly similar to b. However, in general, similarity reasoning might involve more complex *domains* with multiple objects: for example, Rutherford famously used similarity reasoning between the solar system and the hydrogen atom (see Gentner & Jeziorski (1993) for a discussion and analysis of, what is sometimes called, the Rutherford-analogy). For ease of our discussion, I will focus on *single object domains* – i.e., similarity reasoning involving two (concrete) objects.[1] We call the domain *from* which we wish to extend the *source domain* and the domain *to* which we wish to extend the *target domain.* As I focus on single object domains, I will sometimes use 'source object' as a shorthand for 'object in the source domain' (and similarly for the object in the target domain).

Domains consist of an object (which, in the case of complex domains, may itself consist of multiple objects) and their properties. Of all of these properties, some are known to be shared by the objects in the source and target domains; some are known to *not* be shared; and of some it is unknown whether they are shared. We call these sets of properties, respectively, the *positive* analogy, the *negative* analogy, and the *neutral* analogy. The focus of a similarity argument – i.e., the property of the source domain that we are interested in with respect to the target domain – is a subset of the neutral analogy and is called the *hypothetical* analogy (Bartha, 2010, 2019).[2]

We can now say that a similarity judgement helps us to conclude (from a similarity argument) that a particular property holds of the object in the target domain because of some *known* shared properties with the source object, despite some known properties that differ. Importantly, this is an *epistemic* characterisation in the sense that we focus on those properties that are *known* to be shared, not those that are as a matter of fact shared. Let us consider an example, adapted from Roca-Royes (2017), to make things a bit clearer and to relate the recently introduced terminology to the more general (GSA).

> My table, Messy, can support my laptop (i.e., when I place my laptop on Messy's surface, it doesn't fall through it). Twin Messy is the table of my colleague. Messy and Twin Messy are both rectangular, both composed of atoms, both are solid, and both are in the same office. Messy is white, yet Twin Messy is black. I have named Messy 'Messy', I don't know if Twin Messy is named. I am curious whether Twin Messy can support my laptop.

[1] Some examples in this chapter may involve domains with multiple objects, but it will often be easy to see how these examples relate to single object domains.

[2] Note that in the case of complex domains with multiple objects, it is *not* the case that we are interested in reasoning of the form: 'all objects in the source domain have property P, every object in the source domain is relevantly similar to every object in the target domain, thus all objects in the target domain have property P'. Similarly for the notions of positive, negative, and neutral analogy. I take it that we should view the objects that constitute complex domains to be part of a system or complex object which is the object of the source domain (if phrased in terms of a single object). So a positive analogy of similarity reasoning involving complex domains could be: 'there is something in the source domain that has property P' and there is a corresponding object in the target domain that has property P.

The properties 'being-rectangular', 'being-composed-of-atoms', 'being-solid', and 'being-in-office-F2.08' are known to be shared by Messy and Twin Messy – i.e., they constitute the positive analogy. Conversely, of the properties 'being-white' and 'being-black' it is known that they are not shared – i.e., they make up the negative analogy. Of two properties, in this toy example, it is unknown whether they are shared, 'being-able-to-support-laptop' and 'being-named'. These are the neutral analogy and a subset of the neutral analogy – the property I am interested in – is the hypothetical analogy, in this case the property 'being-able-to-support-laptop'.

Note that this is just a systematic way of describing similarity reasoning according to (GSA): we find that Messy and Twin Messy share a number of properties and on the basis of this we might conclude that they also share a further property, the hypothetical analogy. However, at this point we are not yet in a position to judge whether or not concluding that Twin Messy *can* support my laptop constitutes good or bad similarity reasoning. We still haven't said anything about what constitutes *relevant* similarity. To do so, we first need to discuss the *vertical relations.*

7.1.2 Vertical Relations

The positive, negative, and neutral analogy all concern the properties of the object in the domain (in particular, they focus on whether it is known that the object in the target domain also has the properties of the object in the source domain). However, what we haven't considered so far is the *relation between the properties* of the object in the source domain (and, correspondingly, of the object in the target domain). We call these relations between the properties of the object in a domain the *vertical relations.* Note that the vertical relations we are interested in are always *with regards to the hypothetical analogy.* So, if we are interested in whether Twin Messy can support my laptop, then the vertical relations are the relations between (some) properties of Messy and the property of 'being-able-to-support-laptop' – i.e., the hypothetical analogy.

At this point, it will be good to briefly focus on *complex* domains of systems that might themselves consist of multiple objects to get things clear (we will return to single object domains after this discussion). When dealing with complex domains, vertical relations are both two-place relations *between objects* as well as higher-order relations *between properties* (Gentner, 1983; Gentner & Markman, 1997; Bartha, 2010). For example, when we consider a complex domain that consists of objects in my office, then examples of two-place relations between, e.g., my laptop and the table are 'being-supported-by' and 'being-on-top-of'. However, these tell us nothing yet of the higher-order relations between properties of, e.g., my laptop (e.g., the fact that the property 'having-a-full-battery' is related to 'being-able-to-turn-on'). In single object domains, vertical relations are *only* higher-order relations between properties, so this is not much of an issue; it is when one extends this to complex domains that one has to be careful. For example, Bartha (2010), who does not draw the distinction between single object and complex domains, is not always clear on what he means when talking of vertical relations as he takes these to be relations "between the objects, relations, and properties within each domain" (p. 14). Yet, it is important to keep two-place relations between objects and higher-order relations between properties distinct, because, as we will see, it is the latter that are of crucial importance for successful similarity reasoning.

Let me summarise the distinctions and terminology introduced so far. Similarity judgements concern judging two objects to be similar based on the properties that are known to be shared or known not to be shared – i.e., the positive and negative analogy respectively. Similarity reasoning involves projecting the property of interest – i.e., the hypothetical analogy – from the source domain to the target domain based on such similarity judgements. Finally, I suggested that the higher-order relations between the properties of the objects in each domain – i.e., the vertical relations – are of crucial importance.

In the next section, I will specify how vertical relations play an important role in classifying the similarity relation – i.e., the relationship between the source object and the target object that is taken to ground the similarity judgement.

7.2 *Relevant* Similarity

Within the literature on analogical and similarity reasoning, it is common practice to distinguish between *surface similarity* and *predictive analogy* as kinds of similarity relations that result in similarity reasoning of different predictive strength. That is, assuming that the similarity judgement holds, the different similarity relations that give rise to the similarity judgement *affect* the likelihood that the similarity reasoning has a true conclusion.[3] The difference between surface similarity and predictive analogy as similarity relations concerns what vertical relations we take to be important. The surface similarity relation suggests that we can make similarity judgements based on *any* properties shared between the two domains.[4] That is,

[3] Within this literature, people also often distinguish between, what they call, *anomalies* and *literal similarities* (Gentner, 1983; Gentner & Markman, 1997). The former are instances where the source object and the target object share no properties whatsoever; similarity reasoning based on anomalies obviously fails to be proper justification for its conclusion (even if it happens to be true). It is less clear that the latter is supposed to be. Sometimes, these seem to be cases where the source object *is* the target object (Gentner & Markman, 1997, p. 48, Figure 1), whereas other times it seems to involve sharing a high number of vertical relations as well as first-order properties. If literal similarity is supposed to be identity, then similarity reasoning based on it does not extend one's knowledge. If, on the other hand, it is supposed to be something weaker, then I take it that the crucial reasoning step hinges on the shared vertical relations involved (as will be discussed in the next section).

[4] The term 'surface similarity' might be misleading in that it suggests the focus on 'surface' or 'observable' relations, which is not the case.

surface similarity suggests that the relevant vertical relation is that of *mere co-instantiation*.

In this section, I will discuss the surface similarity relation when used as the similarity reasoning involved in similarity-based epistemologies of possibility. In particular, I will argue that the surface similarity relation, in general, does not result in successful similarity reasoning as it fails to take into account the aspect of *relevance*. In the next section, we will discuss the predictive analogy similarity relation.

7.2.1 Surface Similarities

When we use similarity reasoning, we are interested in finding out whether the hypothetical analogy holds of the target object. That is, whether the object in the target domain has the property of interest. Using surface similarity as the similarity relation for the crucial similarity judgement suggests that *any* kind of shared properties should be taken into account when engaging in similarity reasoning. So, the higher-order relations between properties of the object in the source domain (i.e., the vertical relations) are mere co-instantiation (I will designate the co-instantiation of two properties with 'AND(P, Q)'). So, if I am currently listening to music and I am currently wearing a grey shirt, then there is a higher-order relation of co-instantiation between these two properties – i.e., 'AND(listening-to-music, wearing-grey-shirt)'. Suggesting that co-instantiation is the crucial higher-order relation between properties of the object in the source domain and the hypothetical analogy (a particular property of the source domain), is the same as suggesting that we take into consideration *all* similarities. That is, there is no notion of *relevance*.

To see the effects of similarity reasoning based on surface similarity in the epistemology of possibility, let us look at an example of similarity reasoning that Roca-Royes (2017, p. 236) "find[s] epistemically defective:"

> Malala could have had my (human) neighbour's origins (or anyone else's origins). My neighbour had those origins and Malala is not different from [her] in any relevant sense. (ibid.)

Many take it to be *impossible* that Malala has Roca-Royes' neighbour's origins (Kripke, 1980) in which case we would not want our theory to suggest that we are justified in believing it to be possible.[5,6] However, note that with the surface similarity as our basis for the similarity-based epistemology of possibility, it seems that we

[5] We ignore the case where Malala could be living next door to Roca-Royes.

[6] Roca-Royes herself has a milder view in that she thinks that "the current knowability model [should *not*] help elucidate the knowability conditions of" the claim that Malala could have my neighbour's origins, even if it were true (Roca-Royes, 2017, p. 236). On such an account, it would still be problematic if our theory of similarity reasoning suggests that we are justified in believing it to be possible, as Roca-Royes points out.

would be justified in thinking that Malala could have Roca-Royes' neighbour's origins, especially when we make the surface similarity extremely strong. Consider the following shared properties between Malala and Roca-Royes' (hypothetical) neighbour, which *all* instantiate the vertical relation of co-instantiation (the dots indicate that we can extend the list with any number of arbitrary attributes shared):

1. Is a female.
2. Consists of atoms.
3. Is 22 years old.
4. Breathes air.
5. Has a space-time location.
6. Has ten fingers.
7. Is activist.
8. Owns a pair of shoes.
9. Drinks water.
10. ...

Given that Roca-Royes' (hypothetical) neighbour and Malala share all these attributes, we would be justified, on a surface similarity model, to conclude that Malala could also have another property that Roca-Royes' neighbour has, namely being born from certain parents. At best this is not something that we should want our model to predict (even if true) and at worst, this would be a prediction that would be strictly false (if impossible). So, it seems that the surface similarity relation is not a good basis for a similarity-based epistemology of possibility.[7]

I am *not* suggesting that there are no other relations between the properties that Malala has, there are many (e.g., 'being-female' and 'breathing-air' are related in other ways than mere co-instantiation). What we are evaluating here is whether *just* focusing on similarity reasoning based on surface similarity can provide us with justification for accepting the conclusion. The Malala-example is supposed to show that it cannot, as there are many instances where surface similarity-based similarity reasoning predicts the, intuitively, wrong results.

The problem with the surface similarity relation is that it fails to capture any informative structure between the properties in the source domain and the hypothetical analogy; there is no information about what makes the object in the source domain have the hypothetical analogy. That is, there is no sense of *relevance* between the properties for successful similarity reasoning. Without such a notion of relevance, any comparison between domains will seem 'similar' as any two domains share any number of properties (Goodman, 1972; Lewis, 1983; Morreau, 2010).

Researchers who focus on analogical and similarity reasoning have converged on the view that we should focus on *informative* higher-order relations between

[7] See Bartha (2010, p. 197) for another example of similarity reasoning based on the surface similarity relation. The example is of Franklin's reasoning, on the basis of a surface similarity judgement, that lightning would be attracted by metal rods, however, this did not involve a possibility judgement, nor is it obvious that Franklin intended this similarity reasoning to *justify* the conclusion, rather than to 'explore' the idea, to ultimately be justified through experiments.

properties, rather than on mere co-instantiation. The reason that people reject such surface similarity-based similarity reasoning, they point out, is precisely because the surface similarity relation fails to be sensitive to any informative relation between the properties of the object in the source domain and its having the property of the hypothetical analogy (see Hesse, 1966, p. 109; Gentner, 1983, p. 161; Davies, 1988; Russell, 1988; Gentner & Markman, 1997, p. 48; and Bartha, 2010, p. 197).

7.2.2 Vertical Relations as Causal Relations

Besides mere co-instantiation, how do the properties in a domain relate to each other and, in particular, how do they relate to the hypothetical analogy – i.e., the property we are interested in projecting? The reason that we are not interested in mere co-instantiation is that this does not tell us anything *informative* about how the properties of the source object are related. Yet it is such informative relations between the properties of the source object that we are ultimately interested in. Consider the example of whether or not a cup could break. What we need to know for successful similarity reasoning is whether the two cups in question share the properties that, in the broken cup, are related to the property 'breaks'/'is-broken' (e.g., 'being-of-material-X', 'having-forces-Y-acted-upon-it', and the relations between such properties). The question becomes: what *kind* of relations are generally informative in this sense?

Traditionally, philosophers suggested that we should focus on *causal relations*. For example, Hesse notes that we should think of similarity reasoning "as essentially *a transfer of causal relations* between some characters from one side of the analogy relation to the other" (1966, p. 99, emphasis added).[8] Importantly, it seems that Hesse focuses on *direct causation*, which, following Humphreys (1980, p. 309), is non-spurious causation without any events between the cause and the effect, such that the event affects the likelihood of the cause.[9] That is, direct causation is causation without intervening factors. For example, when Hume uses a cue to exert force on a billiard ball, this is the direct cause of that billiard ball moving.

Focusing only on direct causation limits the kinds of cases where we can be said to use similarity reasoning. In a contemporary refinement of Hesse's analysis, Bartha (2010) argues that focussing solely on direct causation "is too restrictive" and that we should "replac[e] [the] causal condition with a more general requirement" (2010, pp. 43-44).[10] In order to account for more structural relations than a

[8] An example of where she is very explicit about this is when she says that "it is not justifiable to pass by analogy from [source] to [target domain] in respect either of properties which are not essential to the [source] or of causal relations of a kind which are not appropriate to [source] or [target]" (Hesse, 1966, p. 98).

[9] Humphreys' definition is a bit stricter. He suggests that there should be no event, B, between the cause, C, and effect, E such that $\mathtt{Pr}(E|BC) = \mathtt{Pr}(E|C)$ (where '$\mathtt{Pr}(\varphi|\psi)$' is the *objective* probability of φ given ψ). See also Woodward (2003, pp. 54-55).

[10] Bartha allows a whole range of relations, including, for example, the relations of 'being a

known direct causation relation, Bartha appeals to Humphreys' (1981) *aleatory explanations.* Aleatory explanations are explanatory relations that are broadly causal and more general than direct causation, also taking into account explanations that rely on a common cause structure or counteracting causes, i.e., causes "which lower the probability of the effect" (Humphreys, 1981, p. 227). For example, when we say that the cup broke because it fell off the table despite landing on carpet, we cannot analyse this *only* relying on direct causation, but we can with aleatory explanations (Humphreys, 1981, p. 227; Bartha, 2010, p. 114). While being more general than direct causation, these aleatory explanations still capture the tacit psychological preference we have "for coherence and causal predictive power" in the similarity relations on which our similarity reasoning is based (Gentner & Markman, 1997, p. 47). When it comes to similarity reasoning, the higher-order relations between properties (i.e., the vertical relations) that we are interested in include, at least, (Humphreys') aleatory explanations.[11]

As theories of similarity reasoning are mainly inspired by and appealed to in (the philosophy of) science, they generally fail to take into consideration other 'metaphysical explanatory' relations such as grounding relations, essential relations, mereological relations, et cetera. Given that we are interested in the epistemology of modality, we should keep open the possibility that these kinds of metaphysical relations also play an important role in similarity-based possibility judgements. In order to include all of these, as well as the aleatory explanations, one may call the relations that we are interested in more generally *structural* relations; allowing one to remain relatively agnostic about exactly what these relations are. Even though in theory I remain agnostic about the exact nature of these structural relations, in practice, in order to simplify the discussion, I will focus exclusively on shared aleatory relations – e.g., causal relations extending beyond direct causes to include common causes, causal chains, counteracting causes, etc. – in this chapter and I will use 'causal relations' to denote this broad class.[12] I do so because of two reasons.

First of all, we are concerned with providing a cognitively plausible epistemology of possibility that explains how ordinary people gain knowledge of mundane possibilities (e.g., this coffee cup could break). If we think that in gaining knowledge of such mundane situations we rely on anything like these structural relations in ordinary life, it is unlikely that these involve essential or grounding relations. As

proof for', to also account for mathematical analogies. Given our focus on the epistemology of possibility, we limit ourselves to predictive analogies, involving aleatory relations (see below).

[11] Moreover, most influential theorists on analogical and similarity reasoning think that there are no serious competitors to these causal relations when it comes to what makes ordinary similarity reasoning successful.

[12] Let me stress that this class includes relations such as 'the glass broke (partly) because it is made out of material X'. So, similarity with regards to 'being-of-material-X' would be relevant if we are interested to see if another glass could also break because of the structural relation between 'being-of-material-X' and 'being-broken'.

Roca-Royes puts it, "[w]e know that my office wooden table can break; [but] it's not so clear that we know that (whether?) its material origins are essential to it—even less so to which degree, if they are (known to be) essential" (2017, p. 223; Hawke, 2017 makes similar remarks).[13] We do, however, seem to rely on causal relations much more often and are reasonably reliable at reasoning on the basis of such causal relations (Strevens, 2000; Gelman, 2003; Nichols, 2006a; Hayes & Thompson, 2007; Cimpian & Salomon, 2014).[14] So, causal reasoning more plausibly has a place in an explanation of everyday modal judgements than reliance on more metaphysical explanatory relations such as grounding or essences.

Secondly, these further structural relations are *modally stronger* than causal relations. That is, relations such as grounding, essence, and material constitution 'go beyond' mere causal modality (whatever that may be) in that they all are closely related to pure metaphysical necessity (see for example, respectively, Bliss & Trogdon, 2016, §5; Fine, 1994; Kripke, 1980). What I mean by 'modally stronger' is that if something grounds, constitutes, or is the essence of something else, then the relation between these two objects needs to hold in more worlds (namely all metaphysically possible worlds) than when the relation between two objects is that of cause and effect.[15] We would not only need to know these relations, but also their modal status in order to justifiably conclude something from the resulting similarity reasoning.[16] The problems that we will raise for similarity reasoning based on prior explicit causal knowledge concern the modal profile of causation in relation to knowledge of everyday possibilities that it is supposed to be epistemically prior to. These worries will all carry over to these (modally) stronger structural relations.

A final terminological note, I will sometimes talk of '*relevant* causal relations'. By this I mean the causal relations that are relevant for the hypothetical analogy (i.e., the property we are looking to project). So, if there are causal relations that are known to be completely irrelevant for the hypothetical analogy, then these causal relations are irrelevant.

7.2.3 Vertical Relations as Relevance

The role of the vertical relation is absolutely critical and is related to the problem of relevance in similarity reasoning. As we have said a number of times before, any two objects will have countless properties in common, so if we do not constrain the

[13] Remember also our discussion of and, in particular, the motivation for, focusing on an epistemology of possibility, rather than necessity from Chapter 1 (Section 1.2.2). It seems plausible that children know that the glass window could break, but it is not obvious that they know the essential or necessary properties of the window.

[14] We will see empirical evidence of everyday causal reasoning from humans, as well as some further subtleties, in Chapter 8.

[15] Though as we will see in Section 7.4, this depends on one's interpretation of causation.

[16] This is needed because we need to be able to distinguish these relations from mere co-instantiations, we need to know that they are in fact structural relations between properties.

similarity relation by relevance, any two objects count as similar (e.g., Aronson et al., 1995, p. 21). The higher-order relations between the properties of the object in a domain allow us to focus on a subset of the properties of that object. Determining what kind of higher-order properties one focuses on is a way of rephrasing this problem of relevance. Taking mere co-instantiation as the vertical relation that determines proper similarity relations (and thus proper similarity reasoning) falls short of this task, as it comes down to, again, having no constraints on which properties are relevant for similarity judgements.

We saw above that a proper similarity relation is one that takes into account the properties that are structurally/causally related to the hypothetical analogy. Similarity judgements based on such a similarity relation make for good similarity reasoning. Phrasing things in terms of *relevant* similarity, this suggests that causal relations (with regards to the hypothetical analogy) determine relevance. So, relevant similarity is similarity in terms of the properties that bear a causal relation to the hypothetical analogy.[17] Let me also stress that these vertical relations need to be *known* in order for the conclusion of the similarity reasoning to be justified. If not, then the agent will not be aware if they are reasoning based on the surface similarity or predictive analogy similarity relation; thus, for all they know, the conclusion may not be justified (if it turns out that there are no causal relations). Successful similarity reasoning involves knowing that the similarity judgement concerns a *relevant* similarity, which in turn depends on knowing that the two objects involved share the properties that in the source object are causally related to the hypothetical analogy.

7.3 Predictive Analogies and Causal Knowledge

Similarity reasoning based on similarity relations with causal higher-order relations between the positive, negative, and hypothetical analogy are called *predictive analogies* (Bartha, 2010). Bartha presents the most thoroughly developed philosophical account of predictive analogies, the *Articulation Account*. In this section I will discuss this theory and argue that, in order for such similarity reasoning to be successful, it requires *prior* justified beliefs (or knowledge) of explicit causal relations. Secondly, I will briefly fend off a potential objection, namely that *syntactic accounts* of predictive analogies provide a counterexample to the claim that reasoning by predictive analogy requires prior causal knowledge. I will argue that they do and conclude by generalising to the conclusion that predictive analogical reasoning in general requires prior explicit causal knowledge. In the next section, I will evaluate how this reliance on prior structural knowledge affects similarity-based epistemologies of possibility that rely on predictive analogies – i.e., *predictive analogy-based similarity theories.*

Let me stress that the conclusion – that analogies require prior explicit causal

[17] As noted in Chapter 6 (Section 6.1), this is in a sense what Hawke (2011) suggests.

knowledge – is *not* to be taken as an objection against these theories of analogy *as theories of analogy*. The arguments that follow in this chapter concern *epistemologists of possibility* who based their similarity theories on such predictive analogical reasoning.

7.3.1 Articulation Accounts of Predictive Analogy

Bartha (2010) develops the *Articulation Account* of analogies and analogical reasoning, which he himself takes to be "a refinement" (p. 35) of the classical account of Hesse (1966).[18] On Bartha's account there are two crucial features of a predictive analogy: *prior association* and the *potential for generalisation*. The prior association are all those properties that are *known* to be relevant ('critical' in Bartha's terminology) for the hypothetical analogy in the source domain. This is where Bartha relies on Humphreys' aleatory explanations discussed above and notes that prior association is the determination of the source domain having the hypothetical analogy because of certain other properties and despite some other properties (2010, p. 114, Definition 4.5.1). The potential for generalisation, roughly, states that there should be some of the relevant properties in the target domain, but not any known defeaters. That is, some positive causal factors are in the positive analogy; none of the known positive causal factors should be in the negative analogy; and none of the known defeaters (of the hypothetical analogy) should be in the positive analogy (idem, pp. 117-118, Definition 4.5.3).

Epistemically speaking, the articulation model requires the following two step procedure:[19]

1. "*Determine relevance (critical and secondary features).* [...] [S]ort out which features of the source and target domains are relevant to the conclusion of the argument, and [...] determine their degree of relevance" (Bartha, 2010, p. 102, original emphasis).

2. "*Assess the potential for generalization (plausibility screening).* The prospects for generalizing the prior association are evaluated by assessing both positive and negative evidence" (Bartha, 2010, p. 103, original emphasis).

We can say that in order to be justified in believing the conclusion of predictive analogy-based similarity reasoning, we have to determine which properties are *critical* (i.e., relevant) and see if some of these are in the positive analogy, while none of

[18] The articulation account is very subtle and complex and I can only give a rough, informal overview here (for a complete account see Bartha, 2010, especially chapter 4). In particular, Bartha's model accounts for a whole range of similarity relations *other than* predictive analogies. For example, his model is able to account for mathematical analogies (Ch. 5) as well as 'weaker' similarity relations such as, what he calls, 'correlative analogies' (Ch. 4.9 & 6.2). I will leave both of these aside.

[19] In the original, Bartha has an additional step that precedes these two: paraphrasing the 'prior association' into a particular canonical form. However, for our purposes this step is not important.

them are in the negative analogy. So, determining what the critical properties are is essential on this account of predictive analogies. According to Bartha (2010, p. 116, first emphasis added) "[a]ll identified *contributing causal factors*" are critical and "[a]ll *salient* defeating conditions [...] for these contributing causal factors are critical (that is, their *absence* is critical)." This means that we need to be justified in believing, among other things, what the contributing causal factors are to the having of the property of interest (i.e., the hypothetical analogy).

So, on Bartha's articulation account of analogies – as our discussion of it, in particular the last quote from Bartha, shows – successful predictive analogical reasoning requires prior explicit causal knowledge (for those analogies we are interested in, e.g., ignoring mathematical analogies).

7.3.2 Formal Accounts of Analogy

Formal accounts of analogies are originated in the computational sciences and aim to provide a purely *syntactical* analysis of analogies (e.g., Gentner, 1983; Falkenhainer et al., 1990; Gentner & Markman, 1997; Forbus et al., 1998). Besides discussing this theory as a feature of interdisciplinary completeness (e.g., acknowledging theories of analogy outside of the field of philosophy), these theories are of particular interest as they seem to *not* rely on prior causal knowledge. That is, they seem to provide a counterexample to my suggestion that predictive analogical reasoning requires prior causal knowledge. The reason one might think is that these accounts promise a completely syntactic analysis of predictive analogical reasoning, seemingly without looking at the content of the positive, negative, and hypothetical analogies. I will argue that even on the analysis of these theories, predictive analogical reasoning requires prior causal knowledge.

The structure-mapping theory, one of the most well-known computational models of analogies (Chalmers et al., 1992, p. 205), exploits the distinction between properties and higher-order relations between these properties. Based on this distinction, structure-mapping theorists suggest a particular schema for finding successful analogies (Gentner, 1983, p. 158). The first two steps of the schema merely suggest to ignore ordinary property sharing and focus on higher-order relation sharing. The third principle, the Systematicity Principle (henceforth: SP), does a lot of work for these formal accounts, so it is worth pausing at it for a moment. The question that SP is supposed to address is: how do we know *which* higher-order relations to take into consideration and which to ignore? That is, what are the *relevant* properties? Formal accounts come up with an answer that, importantly, is supposedly *independent* of our knowledge of the particular objects involved. The systematicity principle suggests that we look at *the largest system of properties* that are related to each other by higher-order relations (Gentner, 1983, pp. 158-164).

The reliance on prior causal knowledge is obscured by the distinction between the *AI system*, which tries to determine the best analogy mapping, and the *domain-*

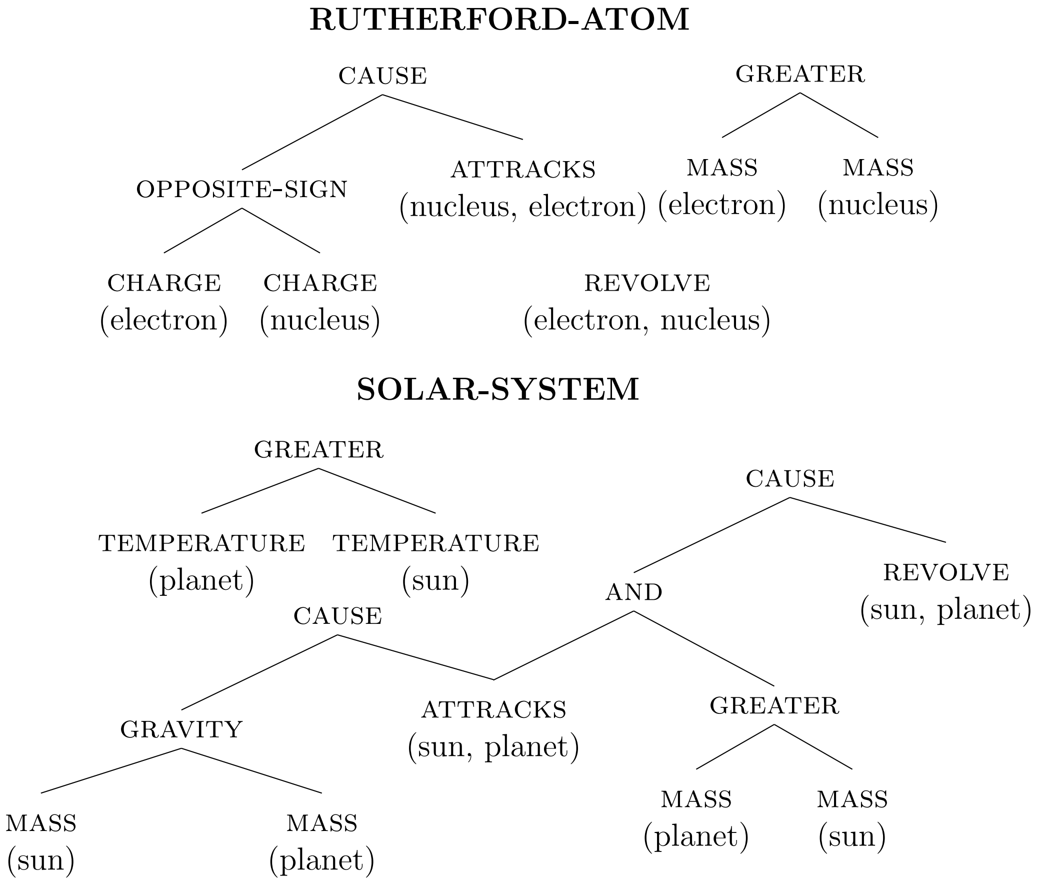


Figure 7.1: *The representation used by the formal account of the analogy between the atom and the solar system (adapted from Falkenhainer et al., 1990)*

expert, who hand-codes in all the relevant relations.[20] The AI system does not look at the content of analogy, but only looks at the relational structures that it is presented with in order to apply the Systematicity Principle. However, here is the crucial part: the saliency of relations is given by a human programmer or subject who *does* know the relevant contents. Our interest here concerns *human* similarity reasoners and, as formal account theorists have admitted, for these the representation is important (Forbus et al., 1998, p. 253). We, as similarity reasoners, are, in a sense, both the AI system and the domain-expert: we first need to decide which relations we take into account and then consider which ones we map onto the target domain. The fact is that this first step – the one that mimics the role of the domain-expert – relies on prior causal knowledge. Consider the Rutherford-analogy between the solar system and an atom: when representing this similarity judgement, we need

[20] Thanks to Arianna Betti for making me aware of this distinction and the implications it has for my argument.

to decide *which* co-instantiated properties are relevant and which ones are not (see Figure 7.1, which only shows the *mapped* relations). We do so based on *knowing* which of these co-instantiated properties play a role in further (causal) relations. This requires prior knowledge of what the relevant causal relations are.

7.3.3 Redundancy, Determination, and Causal Knowledge

One of the problems raised for the Aristotelian analysis of reasoning by similarity, was that it involved a premise of the form '$\forall x(P(x) \Rightarrow Q(x))$'.[21] Many have pointed out that even though including such a premise may solve the problem of justifying conclusions from similarity arguments, it does so by trivialising the similarity reasoning. For example, Davies points out that

> the condition of the similarity P being relevant to the conclusion Q needs to be weaker than the inheritance rule $\forall x(P(x) \Rightarrow Q(x))$, for then the conclusion in plausible analogies would always follow just by application of the rule to the target. Inspection of the source would then be *redundant*. (1988, p. 231, emphasis added)

He defines this as the *Nonredundancy Problem*: the source domain should not be made redundant in an account of similarity reasoning. One may worry that suggesting that similarity reasoning relies on causal vertical relations *also* results in redundancy. The reasoning goes as follows: I know that properties P and R are causally related to property Q and I know that object y has properties P and R, thus, I conclude, that object y has property Q. This kind of reasoning would make any source object irrelevant (the source object in this case would be *relevantly* similar in having the properties causally related to Q in common with y).[22] Let me briefly dispel the worry that suggesting that predictive analogical reasoning – i.e., similarity reasoning based on a similarity relation that takes the vertical relations to be causal relations between properties – requires prior causal knowledge, makes the source object completely irrelevant.

The kind of prior causal knowledge we need to justify the conclusion of predictive analogical reasoning can be phrased in such a way that it does not make the source object irrelevant. The way to do so, is to phrase things in terms of, what Davies (1988), calls, *determination rules*.[23] These are rules that tell us certain abstractions of properties are causally related to each other, without specifying the particular instances of these relations (which would make the source redundant). An example from Davies helps to explain things. Consider Sam and Blake, who both own a

[21] See Hesse (1966, ch. 4); Bartha (2010, ch. 2.2); and Bartha (2019, §3.2) for references to Aristotle's original work on arguments from likeliness and for the full Aristotelian analysis of which the syllogistic inference is only a part.

[22] Thanks to Peter Hawke for pushing me on the redundancy problem for the causal relations.

[23] Russell (1988, p. 257) calls these determination rules "causal factors".

second-hand fire truck. Both of their trucks were built in 1986 and are of the brand Cadillac. Assume that we know that Sam's truck is white and that it is worth €500. Taking the causal relations to be what determines relevance, we can explain that it would be *bad* similarity reasoning to conclude that Blake's truck would also be white, but it would be *good* similarity reasoning to conclude that Blake's truck would also be worth, roughly, €500. However, as Davies points out, we do not expect Blake to come to believe that her truck is worth €500 because she has prior knowledge that being a 1986 Cadillac fire truck causes it to be worth €500; if she did, we would not need Sam's truck as a source object for the similarity reasoning. The prior knowledge Blake has, Davies suggests, is of the form "the make, model, design, engine-type, condition and year of a car determine its trade-in value" (1988, p. 233). This does not make Sam's truck (i.e., the source object) redundant, as we need it to justifiably believe a conclusion of a particular instantiation of this causal relation (in this case, that of a 1986 Cadillac fire truck and the value of €500).

Predictive analogical reasoning thus requires prior causal knowledge, without thereby making the source domain irrelevant. The causal knowledge required are in the form of aleatory explanations (Humphreys, 1981) or determination rules (Davies, 1988), which captures that the kind of property of which the hypothetical analogy is an instance (e.g., the price of a second-hand truck) is caused or determined by other kinds of properties, some instances of which are in the positive/negative analogy (e.g., year, brand, model, etc.). We need knowledge of the source object to draw conclusions about particular instances of these causal relations. This means that "we must bring a good deal of *prior knowledge* to the situation to tell us whether the conclusions we might draw are justified" (Davies, 1988, p. 228, emphasis added). In particular, for predictive analogical reasoning, we need to explicitly know *what* the exact structural relations are that are relevant to the hypothetical analogy.

7.4 Causal Knowledge and Similarity Theories

Similarity-based epistemologies of possibility suggest that we gain knowledge of what is possible through similarity reasoning. The similarity reasoning they rely on is (roughly) of the form (repeated from page 123):[24]

Similarity Argument (SA):

P1. x has property P.

C1. x has property $\Diamond P$. (actuality principle)

[24] I am assuming that even though the hypothetical analogy is, strictly speaking, $\Diamond P$, when we engage in this kind of similarity reasoning, we are interested in the properties that have a causal relation to P. If the target object has these properties, then we know that it *could have (had)* P as well – i.e., we know that $\Diamond P$ is true for the target object. Moreover, remember that I use '$\Diamond P$' as sloppy notation for $\lambda z.\Diamond P(z)$ (see footnote 13 of the previous chapter).

P2. x and y are relevantly similar relative to property P.

C2. y has property $\Diamond P$. (from C1 and P2)

So far, what I've argued in this chapter is that one way for **C2** to be justified, is that **P2** concerns the predictive analogy similarity relation – i.e., a similarity relation that involves vertical relations as causal relations (broadly understood). Consider the following toy-example to make this a bit more explicit. I might come to believe that this cup, a, can break based on the fact that a different cup, b, did break. In order for this belief to be justified through similarity reasoning, I need to justifiably believe what properties the broken cup has that *made it so* that it could break (e.g., the particular physiological make-up of the cup, the particular forces that act upon it) and that the cup that hasn't yet broken also has these properties.

If this is all correct, then being justified on the basis of successful similarity reasoning requires *prior* justified beliefs in the (explicit) causal relations – the relevant causal knowledge also needs to be justified, in order to transmit this justification to the possibility claim that we are interested in (Moretti & Piazza, 2018). This is of itself already a significant finding, for it means that if the similarity theorists aim to address the central question of the epistemology of possibility – i.e., how do we *ultimately* acquire knowledge of possibilities – their account hinges on their explanation of how we are justified in believing these crucial causal relations. Something that isn't often explicitly acknowledged.[25]

The question that arises is what the consequences of this are for similarity-based epistemologies of possibility. The answer, it turns out, depends quite a bit on what one takes causation to be – e.g., if you take causal relations to be relations of necessitation, then you need to know that certain properties necessitate another property. So, how one interprets causal relations has a direct effect on what one takes the vertical relations to be that make a similarity relation capture relevance.

In this section, I will discuss three of the main theories of causation and raise some epistemological worries for epistemologists of possibility that rely on such prior causal knowledge.[26] The relation between cause and effect differs in modal profile

[25] Hawke (2011) and Roca-Royes (2017) both hint at similarity having something to do with the causal or intrinsic nature of the objects involved, whereas Hartl (2016) suggests that this is arbitrary.

[26] One omission in the theories of causation that I will discuss is *interventionalism*, or the *structural equations theory*, with regards to causation (e.g., Pearl, 2000; Woodward, 2003; Beebee et al., 2009, ch. 11). Many categorise such theories as sophisticated counterfactual theories (Paul, 2009; Menzies & Beebee, 2019), in which case the arguments of Section 7.4.2 carry over. Moreover, it is not clear that these tell us what causation is, rather than what correct causal inferences are (e.g., Pearl, 2000). That is, these theories are not so much in need of an epistemology, rather they seem to presuppose an epistemology of causation.

Another omission is primitivism, the view that causation cannot be analysed in further, more primitive, notions (Carroll, 2009). As many have pointed out, one of the main issues with primitivism, in general, concerns our knowledge of causation (see Carroll, 2009; Mumford & Anjum, with different accounts of causation. For example, necessitists take there to be a necessary relation between the cause and its effect, whereas regularity theorists deny any modal relation whatsoever. Both raise potential issues for similarity-based epistemologists of possibility as we shall see.

Let me stress that a lot more can be said about each of the options that I suggest for similarity-based epistemologies of possibility, which I can only treat briefly here. Unfortunately, providing a full discussion of each of them is outside of the scope of this dissertation. What these brief discussions intend to achieve is to bring the challenges of a similarity-based epistemology to the foreground and raise them as must-do tasks for theorists aiming to defend such a theory.[27]

7.4.1 Necessitist Similarity Theory

Necessitists think that causation is itself a *necessary* relation (Mumford & Anjum, 2013, ch. 4). For example, Hesse suggests a way of characterising causation that is explicitly modal: "[A] cause A may be interpreted modally, as in some sense *necessary* for B" (1966, p. 79, original emphasis). Similarly, one of the more prominent analyses of causation, i.e., Mackie's (1980), suggests that a cause "is an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" for its effect (p. 60, original emphases). As Mumford & Anjum (2013, p. 43) point out, suggesting that a cause is *sufficient* for its effect is a way of saying that the cause necessitates the effect (see also Carroll & Markosian, 2010, p. 25). In general, on necessitist accounts, causal relations are analysed in terms of necessity. For example, 'A causes B' is true if and only if in some class of worlds, C, if A is true, then so is B. The theorists discussed above take the class of worlds in this case to be *all worlds* (so C is a trivial restriction on the set of all worlds). In terms of the similarity relation, this means that two objects are *relevantly* similar if they have the properties in common that in the source object necessitate having the hypothetical analogy.

Even though these formulations of neccesitism might not be very popular, there is a family of theories that is gaining popularity and that explicitly takes causation to be a relation of necessity: the *powers analysis* of causation (Mumford, 2009). These theories reject the Humean metaphysics and take causation to be a necessary relation between a cause and its manifested effect in terms of *powers* (Mumford & Anjum, 2011). They even argue that anything other than accepting causation to be a necessary relation would leave much of what causation is a mystery (Mumford, 2009).

2013, ch. 8; Schaffer, 2016). It seems very intuitive that what we (perceptually) experience is a mere sequence of events, but then if causation is not analysed in any other terms the question rises: how do we get knowledge of (this primitive notion of) causation if all we see is a sequence of events? Though there are possible responses on behalf of the anti-reductionist, the issue is far from being settled. This is particularly pressing for similarity theorists, whose epistemology of possibility, on a predictive analogy-based account, relies on the possibility of having prior causal knowledge.

[27] I count myself as being one of those theorists and propose a positive story in the next chapter.

Knowing causal relations on such accounts would amount to knowing relations of necessity.

The reliance on prior knowledge of necessity when providing an epistemology of possibility is problematic (as we saw before in Chapter 1 and Chapter 3). If the similarity theorist holds that there is a *uniform* epistemology of modality that is possibility-first, then they would claim that all modal knowledge comes from (or is derivable from) knowledge of possibility. This is particularly plausible for our modal knowledge concerning concrete objects. Remember the example about Roca-Royes' (2017) table, mentioned above: it is intuitively very hard to know necessities about concrete objects (e.g., essences), but is seems intuitively easy to know possibilities concerning them. So it's *prima facie* more plausible that we should focus on an epistemology of possibility, rather than grounding the knowledge of possibilities in knowledge of, modally stronger, necessities. But, as many have pointed out, when providing an epistemology of possibility, the entire project would be undermined if it relied on prior knowledge of necessities (e.g., Hale, 2003; Hill, 2006; Roca-Royes, 2011a,b; Fischer, 2016a; Vaidya, 2016; Roca-Royes, 2017).

The similarity theorists might also hold that the epistemology of modality is *non-uniform*: it might be that there are different approaches – varying in focusing knowledge of possibilities or necessities – for gaining different kinds of modal knowledge. For example, as we saw before, Roca-Royes explicitly advocates a non-uniform epistemology of modality (see Roca-Royes, 2017, 2019b). She holds that we should accept a possibility-first, similarity-based approach for modal claims concerning concrete objects, whereas she suggests a necessity-first (or, more appropriately, an essence-first) approach for our modal knowledge concerning abstracta. Yet even if one opts for such a non-uniform approach, *within* one particular class of modal claims (e.g., concerning concrete objects), the epistemology thereof is uniform.[28] If this were not the case, we would not be able to provide a systematic explanation of how we gain modal knowledge.

So, whether or not one believes that the epistemology of modality is uniform or not, reliance on prior knowledge of necessity within the epistemology of possibility for concrete objects is always problematic. This suggests that, as far as the similarity theorists want to base their account on predictive analogies, there are severe worries for those who also accept a necessitist account of causation.[29]

[28] This was also elaborately discussed in Chapter 1 (Section 1.2.2).

[29] It is an interesting question whether necessitists are committed to take the necessity involved to be *metaphysical* necessity, or whether a weaker notion of necessity may be involved. Taking the causal relation to be one of a necessity weaker than metaphysical necessity perhaps leaves some room for similarity-based epistemologists of possibility to avoid some of these arguments.

7.4.2 Counterfactual Similarity Theory

A modally weaker analysis of causation is the counterfactual analysis. These theories suggest that causation (or causal dependence) can be analysed in terms of counterfactual conditionals (or counterfactual dependence) (Lewis, 1973a; Collins et al., 2004; Paul, 2009; Menzies & Beebee, 2019). Roughly, an effect, e, causally depends on its cause, c, if and only if, if c were to occur, e would occur and when c were not to occur, e would not occur (Menzies & Beebee, 2019, paraphrased from §1.1). Counterfactual dependence, or counterfactual conditionals, are often evaluated as *variably strict conditionals*. This is a *restricted necessity* analysis of causation: in *all* the closest or nearby possible worlds where the cause is true, the effect is also true. In terms of the similarity relation, this means that two objects are *relevantly* similar if they have the properties in common that in the source object necessitate the having of the hypothetical analogy *in all the closest worlds*.

This means that in order to justifiably make similarity judgements, we need to have prior knowledge of counterfactual dependences. If this is analysed in terms of knowledge of restricted necessities, then similar worries as those for the necessitists arise. One of the main epistemologies of counterfactuals, however, suggests we gain knowledge of counterfactuals through engaging with certain imaginative episodes (Williamson, 2005, 2007).[30] However, Roca-Royes (2011b) and Tahko (2012) argue that Williamson's (2007) epistemology of counterfactual conditionals also relies on prior modal knowledge.

> The distinctive feature of [Williamson's epistemology] is that it requires us to hold fixed *constitutive facts*. Furthermore, for our counterfactual judgements to amount to counterfactual *knowledge*, it is not enough that we merely happen to hold fixed the right things – our counterfactual judgements would be (extensionally) correct in this case, but hardly knowledge. We need to hold them fixed *knowledgeably*. This seems to require knowledge of what the constitutive facts are. [...] If this is so, [Williamson's epistemology] implies that [for proper] counterfactual evaluation **we must have prior modal knowledge**. This prior modal knowledge would be a *pre-condition* for counterfactual knowledge.
>
> (Roca-Royes, 2011b, pp. 548-549, original emphases, boldface added)

So, in general it seems that we need to have prior knowledge of restricted necessities or constitutive facts in order to evaluate counterfactual conditionals. Even though weaker than necessity *tout court*, worries arise whether we are basing simple everyday knowledge of possibility claims on modally stronger knowledge of counterfactual conditionals. I should emphasise though that these worries for counterfactual condi-

[30] We elaborately discussed the imagination part of Williamson's epistemology of counterfactuals in Chapter 4 (Section 4.7).

tionals are less pressing than for knowledge of necessities, for it is less obvious that counterfactual knowledge is not a form of mundane, everyday modal knowledge.[31]

Similarity theorists who aim to explain our knowledge of possibility, by relying on our knowledge of causation, which they analyse in terms of counterfactual conditionals, should be careful that the epistemology of counterfactuals does not turn out to be itself reliant on modal knowledge. In particular, if the epistemology of counterfactual conditionals relies on prior modal knowledge, then worries similar to those for the necessitist theories of causation arise.

7.4.3 Regularity Similarity Theory

Theories that explicitly do not rely on any modal relation – i.e., of necessity of counterfactual dependence – between the cause and effect are *regularity theories* of causation (Psillos, 2009). As Psillos puts it, a crucial element of a regularity theory is that "all events of type C (i.e. events that are like c [the cause]) are regularly followed by (or are constantly conjoined with) events of type E (i.e. events like e [the effect])" (2009, p. 131). In the similarity relation, this results in two objects being *relevantly* similar if they share the properties that are the properties that (in the source object) are regularly followed by the hypothetical analogy.

The main epistemological issue for such theories is that "the theory has no resources to distinguish between causes and coincidences. Should there really be no possible distinction between regularities that are genuinely causal and those that are merely accidental?" (Mumford & Anjum, 2013, p. 24).[32] For our purposes, the worry is that because a regularity-theorist cannot distinguish between co-instantiations and genuine causation, we lose the benefits of taking 'relevant similarity' to be the predictive analogy. We can no longer distinguish causal relations from mere co-instantiation on such a regularity theory and, because of this, it is not clear how the regularity-theorist distinguishes between surface similarities and predictive analogies. (This is a problem for any Humean account of predictive analogical reasoning; not just for epistemologists of modality.)

[31] Additionally, one might worry that in order to make possibility judgements based on counterfactuals, we need to also know that the antecedent of the counterfactual conditional is possible (Williamson, 2007; Gregory, 2017). This possibility judgement of the antecedent has to be made independently of and prior to the evaluation of whether the counterfactual conditional is true. This is indeed true, but does not really affect the similarity reasoning based on similarity relations with such counterfactual dependencies as their vertical relations, as we know that the antecedent is possible because the object in the source domain *has* the properties that are causally related to the hypothetical – i.e., the properties that are the antecedent of the causal dependency are actual, thus also possible.

[32] This issue manifests itself differently depending on whether one thinks that we can analyse causes and effects in isolation or whether we should take a God's-eye perspective on all instances of the constant conjunction. In the former case, the issue manifests itself roughly as the problem of induction, whereas in the latter case, one ends up with the paradoxical view that the less often a cause and effect occurs, the easier we can conclude that there is causation.

This is not supposed to be a knock-down argument against regularity theories, but a worry raised for those similarity-based epistemologies of possibility, where the similarity relation has causation as regularity as their vertical relation. One may try to get around these worries by proposing a more sophisticated regularity theory. The worry, however, is that it remains a must-do task for similarity-based epistemologists of possibilities to find a way of dealing with the *inductive vertigo* that regularity theories of causation give rise to – i.e., "that feeling one gets when one spends too long reflecting on the fact that everything [causal reasoning in particular] may yet fall apart at any moment" (Beebee, 2006, p. 532).

Let me summarise. Adopting a weak notion of what we take the relation between cause and effect to be (e.g., a mere constant conjunction relation) results in a very weak form of the predictive analogy, which is no longer obviously predictively stronger than the surface similarity relation. However, the stronger one takes the relation between cause and effect to be (e.g., counterfactual dependence or necessity), the more worries arise for potential reliance on problematic prior modal knowledge. A predictive analogy-based similarity theorist has to walk this balance very carefully. This raises an important challenge for similarity-based epistemologists of possibility that needs to be addressed.

7.4.4 Weak Similarity Theory

Alternatively, similarity-based epistemologists of possibility might opt for, what I will call, a *weak similarity theory.* Weak similarity theorists do not aim to address the central question of epistemology of modality – how can we ultimately come to know what is possible? – but only focus on the *hierarchical question* – "given that there is a distinction between necessity, possibility, and essence, is knowledge of one *more* fundamental than knowledge of the others?" (Vaidya, 2016, original emphasis). For these weak similarity theorists, the findings with respect to predictive analogies seem to be an answer to their question: knowledge of possibility is based on prior causal (structural) knowledge, so, knowledge of causation has epistemic priority over knowledge of possibility (and perhaps modality in general). What I mean by certain sorts of knowledge having epistemic priority is the following. If, in order to have justified beliefs or knowledge of M, I need prior justified beliefs or knowledge of C (i.e., the justified beliefs of C are crucial for the transmission of justification involved in getting knowledge of M), then knowledge of C has epistemic priority over knowledge of M.

Reliance on prior modal knowledge is, in and of itself, not problematic for weak similarity theorists. However, in order for the proposed hierarchy of modal knowledge to be intuitive. Just as our pre-theoretic motivation for focusing on epistemologies of possibility rather than on epistemologies of necessity relied on our intuition that we have many justified beliefs in possibilities *without* having the corresponding beliefs in necessary or essential claims, so too should it be intuitively plausible that the kind of knowledge that weak similarity theorists suggests has epistemic priority over our knowledge of possibility indeed precedes it.

Even though the weak similarity theory is a logically consistent view that is of some interest, I find it unsatisfyingly modest.[33] I think that similarity theorists should not settle for such a modest approach to the epistemology of modality and that they *should* attempt to address the central question of *how* we gain knowledge of possibility.[34]

7.5 Conclusion: Similarity Sweet Spot

In Chapter 6 (Section 6.3), I've spelled out the kind of similarity reasoning similarity-based epistemologies of possibility rely on. In this chapter, I've discussed the general form of these arguments and noted that the crucial premise in such arguments is the similarity judgement. Whether we judge two objects to be similar depends on what we take the similarity relation to be. In particular, we need to specify how this similarity relation is constrained if we want the similarity judgement to carry any justificatory force in the similarity argument. If we do not, any two objects can be said to be similar. In terms of the literature on analogical and similarity reasoning, we need to specify what we take the vertical relations to be in order to determine *relevance* in relevant similarity.

The surface similarity relation takes the vertical relation to be one of mere co-instantiation, which is the same as having *no* restrictions on the similarity relation. We saw that similarity reasoning based on such a similarity relation allows us to justify all sorts of bad instances of similarity reasoning (e.g., accepting that Malala could have someone else's origins). So, looking at the literature on analogical reasoning, we've found that a plausible way of cashing out the notion of 'relevant similarity,' which similarity-based epistemologists of possibility often appeal to, is in terms of predictive analogies.

I have argued that predictive analogies require prior *explicit* causal knowledge, in that the agents need to know the exact relevant causal relations before they can justifiably draw any conclusions from a predictive analogy. How this affects predictive analogy-based similarity theories depends, to some extent, on how they think of causation. For example, one might ignore the precise nature of causal knowledge and focus only on the hierarchical relation between modal knowledge and causal knowledge. These, what I've called, weak similarity theorists content themselves with addressing architectural questions about our knowledge and not

[33] Additionally, given that the theory of causation that weak similarity theorists can rely on has to be modally and epistemologically simpler than that of 'possibility', they might struggle to find a theory of causation that fits the bill.

[34] Similarity theorists like Hawke (2011) and Roca-Royes (2017) do in fact seem to aim to address the *central* question.

the more central question about *how we ultimately gain* modal knowledge. More ambitious predictive analogy-based similarity theorists require prior justified beliefs in causal relations.

In general, relying on predictive analogies requires (ambitious) similarity theorists to spell out their views on causation and its epistemology. If they don't do so, their theory remains explanatorily incomplete and unsatisfactory *as philosophical theories about the justification of our possibility knowledge*. For if they cannot explain how we get justification in the belief concerning the crucial causal relation, they have done little more than explaining one kind of knowledge (modal) with a kind of knowledge (causal) that is in equal need of explanation. The explanatory value of the resulting similarity theories would be incomplete and unsatisfactory as philosophical explanations of the epistemology of possibility.[35]

This leaves similarity theorist with three options:

(i) Focus on the hierarchical question of the epistemology of modality and on evaluating whether causal knowledge grounding possibility knowledge is in fact a plausible hierarchy.

(ii) Accept that predictive analogies are the best way of cashing out 'relevant similarity'. This means that the ambitions of the project are tempered in that the main focus is now shifted to explaining how we get explicit knowledge of the crucial causal relations, potentially suggesting that knowledge of everyday possibilities simply is (or is based on) knowledge of counterfactual- or necessity-relations between causes and effects.[36]

(iii) Suggest an alternative to the predictive analogy as the similarity relation. The alternative should be such that it is predictively stronger than the surface similarity relation, yet does not rely on problematic prior modal knowledge.

In the next chapter, I will develop option (**iii**). I will do so on the basis of *reasoning by kind*. Kinds and kind judgements are closely related to similarity judgements (see, e.g., Quine, 1969). There is ample evidence that we can reliably come to believe things about the causal core of a particular kind – and properties and behaviour that are *due to* this causal core – *without* knowing explicitly which properties make up

[35] This is, I take it, similar to (part of) the sentiment that Roca-Royes expresses concerning the reliance on prior knowledge of essences. She points out that if an epistemology of possibility relies on prior knowledge of essences, "such an epistemology will not have fully-and-satisfactorily elucidated possibility knowledge *until it has satisfactorily elucidated* essentialist knowledge" (2017, p. 223, emphasis added). So, to paraphrase Roca-Royes, as long as "such capacity for [causal] knowledge is left unsatisfactorily explained, [...] this compromises (the satisfactoriness of) the elucidations they provide of our ordinary possibility knowledge" (2017, p. 244).

[36] This is a general lesson for any epistemology of modality that relies on prior causal knowledge. Similar to the lesson Roca-Royes (2011b) and Tahko (2012) draw with respect to Williamson's reliance on *constitutive* facts.

this core (Strevens, 2000; Gelman, 2003; Cimpian & Salomon, 2014). I will develop a similarity-based epistemology of possibility on such kind-judgements, which have the predictive strength of many shared causal relations without the epistemic burden of knowing exactly what these relations are.

Chapter 8

Kinds and the Epistemology of Possibility

[C]ategories serve as building blocks for human thought and behavior

– Medin, 1989

In this chapter, I propose a new similarity-based epistemology of possibility. I do so by developing the notion of 'relevant similarity' based on the metaphysics of kinds and empirical results from the psychology of reasoning. I will use these two components to provide a rational reconstruction of the similarity reasoning that is cognitively plausible and knowledge-conferring. Very roughly, the idea is that once we know that an object of a certain kind has a particular property, we are justified in believing that all members of that kind *could have* that property. So, once we judge two objects to both be cats and one of the two has the property of 'eating-fish', we can conclude that the other cat could have that property.

I will first provide a rational reconstruction of this kind of similarity reasoning. The two main premises are the generalisation to all members of a kind and the classifying to objects as belonging to the same kind. I will discuss the justification for these two premises elaborately in turn. Then, I note a potential equivocation worry between the notion of 'kind' involved in these two premises. I argue that we overcome this worry by reliance on a heuristic and discuss the resulting fallible epistemology of possibility and some of its theoretical virtues. I conclude by discussing the epistemology of modality of Mallozzi (2018a), which is quite similar to the one proposed in this chapter, and argue that her account falls victim to a mismatch worry that my proposal does not succumb to.

8.1 Kinds and Possibility

Let me start by giving a rational reconstruction that we go through when we are trying to determine whether it is possible for an object, a, to have a particular property, P.[1] Let's call it the '$\Diamond$-argument'. Here is an informal description of the $\Diamond$-argument:

> Could a have property P? We know that a is the same kind of thing as b and that b has (had) property P. If there is a thing of a particular kind, K, that has a particular property, then *all* those things can have that property. So, all the things that belong to the kind that b belongs to, could have property P. Hence, a could have property P.[2]

In order to get to the conclusion, three things are crucial: (1) we make the judgement that a and b are of the same kind; (2) we know (or judge) that b has property P; and (3) we know that if one member of a kind has a particular property, then all members of that kind *could have* that property. Formally, the rational reconstruction can be captured by the following argument:[3]

1. $\exists K(Ka \wedge Kb)$
2. Pb
3. $\forall K \forall P'(\exists y(Ky \wedge P'y) \rightarrow \forall x(Kx \rightarrow \Diamond P'x))$
Con. $\Diamond Pa$

The reasoning thus reconstructed is valid – i.e., the conclusion follows from the premises. We start by discussing premise (3), rather than (1), as this is what captures the ampliative aspect of similarity reasoning. Premise (3) follows from some *metaphysical* assumptions about kinds and I will defend these in Section 8.2. Premise (1) is an *epistemological* claim of our ability to classify objects as belonging to a particular kind and I will argue that this premise is justified in Section 8.3. If we can justify these premises, we have a good grounds for thinking that we are justified in believing (some) possibility claims. However, it turns out that there is a potential equivocation worry between premises (1) and (3); in Section 8.4 we turn to discuss this worry and I argue for a particular solution. After this, I turn to the ways in

[1] Remember that, because I focus on rigid reference to the things involved in the possibility judgements (either by naming or demonstratives, e.g., *that* cup), the *de re* and *de dicto* possibilities come down to the same thing. I will sloppily talk of it 'a possibly having property P' and 'it is possible that a has P'.

[2] I talk of 'thing' here instead of 'object' as I want to leave it open that this kind of reasoning also applies to, e.g., events.

[3] A notational remark: I use 'K' to denote first-order properties with a special second-order property, namely the property of being a kind. E.g., the first-order property, 'being-a-cat', has the special second-order property of 'being-a-kind'. A more tedious way would be to introduce a second-order predicate for 'being-a-kind', e.g. $\mathcal{K}$, and then add this each time we use the first-order property, K. For example, premise (1) would then be: $\exists K(\mathcal{K}(K) \wedge Ka \wedge Kb)$.

which the resulting theory is fallible (Section 8.5) and discuss the theoretical virtues of the theory (Section 8.6). I conclude in Section 8.7 by discussing Mallozzi's (2018a) epistemology of modality, which is based on a similar metaphysical perspective on kinds. I suggest that my theory is more promising as it does not succumb to a problem that I raise for Mallozzi's theory.

In the remainder of this section, I will briefly say something about premise (2). This premise is of no particular interest to epistemologists of modality – premises (1) and (3) are – so we can be brief.

Premise (2) concerns ordinary first-order claims such as 'that cat jumps', 'Quinn drinks coffee', 'this cup is black', et cetera. As long as we focus on ordinary, mundane claims, there is nothing special to the epistemology of them.[4] In particular, given that the modality in our argument comes from premise (3), we can focus on *non-modal* instances of first-order claims for our premise (2). This means that the epistemology of such claims is orthogonal to the interests of epistemologists of modality. One can plug in their preferred story about the justification of non-modal, first-order claims. For example, when I see that the cup that I once had did break, we rely on the epistemology of perception for premise (2) (Alston, 1993, 1999; Lyons, 2017). When a trusted source tells me that their cat jumps, we need an epistemology of testimony for premise (2) (Adler, 2017). So, for our purposes, we can forgo a detailed story about the justification of premise (2) in this reconstruction and plug in whatever one takes the correct epistemology of such claims to be.[5]

8.2 Premise (3): Generalisation to Kind-Members

Let us turn to the main premise in the reconstruction:

$$\mathbf{3}.\ \forall K \forall P'(\exists y(Ky \land P'y) \to \forall x(Kx \to \Diamond P'x))$$

That is, *if* there is a member of a kind that has a particular property, e.g., P, then it is possible *for any* member of that kind to have that property. I will defend this premise by appealing to the *metaphysics* of kinds. In particular, I will be relying on a *technical* notion of kindhood. It is important to keep in mind that this technical notion need not be, and likely is not, the same as the intuitive notion of a *natural* or *objective* kind. In order to avoid confusion, let me call the technical notion needed here: *fundamental kind*.

First a terminological note. In philosophical discussions the term 'natural kind' has often occupied the centre stage. However, as Khalidi (2013, pp. 4-5) notes, this

[4] When we consider claims such as 'Philosophical zombies have no phenomenal consciousness', it is less clear what the correct epistemology should be. This is in line with modal modesty; see Chapter 11.

[5] Note that, once again, we won't engage with the radical sceptic who *denies* that we can have any knowledge of such first-order claims.

term might raise some unfortunate connotations. For example, it might suggest that we are only interested in the kinds from natural science, rather than, for example, the social sciences. Mill (1882) used the term 'real kind' instead, which has fewer connotations, but unfortunately this is not widely used. Again others have used 'objective kind'. All these terms raise many interesting questions (e.g., should we only turn to natural sciences to see what kinds there are), but for the purposes of this dissertation, I want to remain neutral (and agnostic) on most of these. When I am *not* talking of fundamental kinds, I will often simply use the unqualified 'kinds,' though I will sometimes follow the traditional usage of 'natural kind' and 'non-natural kind'. I follow Khalidi (2013) in using 'kinds' for things that pertain to the objective world and 'categories' (or 'classifications') for things pertaining to our language or epistemological practices (this will become clear throughout the next few sections).

8.2.1 Simple Causal Theory of Kinds

The justification of premise (3) is compatible with any metaphysics of kinds that accepts these three important (and common) features:

1. Kind-membership is determined by having a *set of core properties.*
2. Members of a kind have that set of core properties by way of *causal necessity.*
3. Kinds form a *hierarchy.*

From these three features, we can define what it is to be a fundamental kind, which allows us to justify premise (3). Let me go over these three features in more detail before I explain what I take fundamental kinds to be.

Core Set of Properties

What makes it that the set of all cats is a natural kind, whereas the set of all white things is not? Most realist theories of kinds suggest that what distinguishes natural from non-natural kinds is that members of a natural kind share a set of core properties.[6] For example, *metaphysical essentialism* holds that there is a set of properties that make up the *essence* of a kind and that all members of a kind have that essence (Putnam, 1973; Kripke, 1980). Similarly, *homeostatic property cluster*

[6] Concerning the metaphysics of kinds, there are, roughly, three traditional contenders: conventionalism, essentialism, and the homeostatic property cluster theory. I take it that the latter two are instances of realism about kinds and I take it that (strong) conventionalism, which holds that what makes a set of things a kind depends on convention and human interest, is the opposite of realism. Additionally, Bird (2018, pp. 1398-1399) points out that, what he calls, weak realism – i.e., that there is an objective, natural division amongst things – seems to follow from methodological naturalism (at least on a particular interpretation of it). I won't consider conventionalism.

theory (HPC) holds that what constitutes being a natural kind is having a set of properties that are related to each other by a homeostatic mechanism (Boyd, 1991, 1999). Finally, the *simple causal theory* holds that kinds are constituted by a set of properties that are causally related to many of the other properties that many kind members share, however, it rejects the idea that this core set of properties needs to be 'tied together' by a homeostatic mechanism (Craver, 2009; Khalidi, 2013, 2018). I will proceed by focusing on the simple causal theory of kinds, as it has the least commitments beyond the three features mentioned above. However, both metaphysical essentialism and HPC would serve us equally well.[7]

Khalidi (2013, 2018) distinguishes between *primary* and *secondary* properties of a kind.[8] The secondary properties are all of the many properties that are associated with a kind. For example, when we consider the kind SILVER,[9] we know that pieces of silver share many properties, e.g., melting point, boiling point, conductivity of sorts, colour, potential chemical combinations, et cetera. The primary properties are those core properties that 'are responsible' for members of a kind to share all these other, secondary, properties. That is, the primary properties *cause*, in a very broad sense of 'cause', many of the secondary properties.[10] In the case of SILVER, it is the property of having atomic number 47 that *causes* (in a suitably broad sense) members of the kind to have (and thus share) many of these other properties (see Mallozzi, 2018a, p. 9 for a detailed discussion of the silver example).

According to the simple causal theory of kinds, kinds are associated with the cluster of primary properties, where a property is in this set if it causes many of the other (secondary) properties of members of that kind.

> Crucially, [. . .], there is a causal link between properties, with one or a few of the properties being causally prior to the others. What characterizes natural kinds is that, even when one or a few properties are central to a kind, there are a number of other properties associated with that kind that are causally related to them. It is this network of properties

[7] In Appendix C, I will briefly discuss some metaphysical theories in a bit more detail. I will provide some initial arguments against these two theories in support of the simple causal theory, however, this discussion should not be taken to be definitive. See Kornblith (1993); Craver (2009); Khalidi (2013, ch. 1); Bird (2018); and Bird & Tobin (2018) for excellent reviews of these positions and some (additional) arguments in favour of and against them.

[8] As Khalidi points out, these primary and secondary properties are *not* supposed to be confused with Lockean primary and secondary qualities.

[9] I will use SMALL CAPS for kinds – e.g., a piece of silver has many properties that it shares with other members of the kind SILVER.

[10] Note that the kind of causation that is involved in the relation between the common core properties of a kind and many of its other properties or behaviour need not be direct causation. As in the previous chapter, we allow for common cause structures, counteracting causes, etc. Actually, we can remain agnostic about what kind of causation is involved, especially given that I will argue that we don't need to know these relations explicitly (see Section 8.4). For those interested, see Khalidi (2013) and Godman et al. (2020) for excellent discussions of the types of causation involved in different kinds of kinds.

> that seems to distinguish natural kinds from non-natural kinds. The causal relations between the properties in the network ensure that natural kinds are projectible and play a central role in inductive inference. (Khalidi, 2013, p. 204).

That is, very roughly, Khalidi's simple causal theory of kinds (see also Keil, 1995; Craver, 2009; Mallozzi, 2018a; and Godman et al., 2020). I follow Khalidi (2018, p. 1384) in referring to this cluster of primary properties as the 'core' or 'causal core' of a kind. Importantly, which properties make up a core of a kind is something that needs to be discovered by science. However, as we will see in Section 8.4, agents do not need to know the exact combination of properties that make up a kind's causal core when reasoning based on kinds.[11]

Causally Necessary Causal Core

So far, we have seen that a kind is associated with a causal core in the sense that being a member of a kind, K, means that you have a certain set of core properties, $C_K = \{C_1, \ldots, C_n\}$, which cause many of the other properties that members of that kind share. Not all secondary properties *need* to be instantiated in all members of a kind (in general, causal relations involve *ceteris paribus* clauses). Additionally, there are many properties that members of a kind can have (e.g., having lost a toe) that are compatible with the causal core of that kind, but that are not caused by it. As Shalkowski (1992) points out, relations such as those between the primary, secondary, and other properties are in the back of our minds when we assert certain counterfactual situations. In particular, there is a class of possible situations that is suitably similar to the actual world that is relevant here; call it *causal modality* (Shalkowski, 1992; Shoemaker, 1998).

Shalkowski takes causal modality to concern the lawful connection and relations between nonaccidental events (1992, p. 56), whereas Shoemaker suggests that it is the modality concerning the special status of "causal laws and their consequences" (1998, p. 59). Lowe takes causal modality to be modality with respect to the "causal laws that actually reign" (2012, p. 919). How to spell out causal modality exactly is a tricky issue, I provide an initial characterisation following the general approach to modalities of Kment (2014): we define causal necessity by defining when worlds are members of a sphere of causally possible worlds around actuality.[12]

[11] Remember, HPC theorists or metaphysical essentialists about kinds also should accept this, but they would argue, respectively, that this core of a kind is 'held together' by a homeostatic mechanism or constitutes the essence of a kind.

[12] We can let 'the actual world' be non-rigid in such a way that there are different spheres around different possible worlds (though we don't need to). If we do this, we can account for the view of, e.g., the simple causal theory and HPC theory that kinds have their core contingently. Others might favour fixing the actual world rigidly.

Definition 7. A proposition is causally necessary iff it is true in all the worlds that have the same causal laws as the actual world.

Causal modality as sameness of causal laws will do for our purposes and we can ultimately adopt whatever the correct definition turns out to be.[13] Moreover, we can leave it open if causal necessity relates to or is the same as natural or nomic modality.

One thing that we do need to assume is something akin to *plenitude* (Lewis, 1986, sec. 1.8). For our purposes, we say that for any object x of a kind K, we assume that for each property P that is compossible with the causal core of K but not a part of it, there is a causally possible world where x has P and a causally possible world where x lacks P in (re)combination with any other properties compossible with the causal core. Call this the *causal plenitude postulate*, 'causal plenitude' for short. For example, being a member of the kind CAT is compossible with the property of being painted yellow. So, if Lou is a cat, then there is a causally possible world where he is painted yellow and one where he isn't.[14] We can make this a bit more precise. The set of properties that make up the causal core of kind K is denoted by C_K. We say that a property P is *merely compossible* with C_K if (i) it is *not* part of the causal core – i.e., $P \notin C_K$ – and (ii) it is compossible with C_K – i.e., $\Diamond(P \wedge C_1 \wedge \cdots \wedge C_n)$. We use '$\mathfrak{C}(P, K)$' to indicate that P is merely compossible with the causal core of K. Then, the minimal condition for causal plenitude is the following:[15]

$$\text{(CP)} \qquad \forall x \forall K \forall P'((\mathfrak{C}(P', K) \wedge K(x)) \rightarrow (\Diamond P'(x) \wedge \Diamond\neg P'(x)))$$

The motivation for causal plenitude can best be appreciated when the full argument for premise (3) is on the table, so I will address it below in Section 8.2.3, under the 'modal scarcity worry'.

Whatever exactly the causally possible worlds are, I take it that it is a feature of being a member of a natural kind that objects have their causal core in *all* causally

[13] See Kment (2014, pp. 184-189) for some subtleties concerning defining modalities in this way. For example, Kment defines modalities in terms of 'matching laws' with the actual world, where this 'matching' can be interpreted in a number of different ways.

[14] Given that causality always involves a *ceteris paribus* clause, there might be causally possible worlds where a member of a kind *lacks* certain secondary properties – i.e., the properties that are caused by the causal core. For example, even though 'being-a-white-shiny-metal' is a secondary property of the members of the kind SILVER, it might be that a piece of silver is coated with something in such a way that it is not white and shiny (this is similar to the issue of, for example, finks for dispositions).

[15] Note that the quantification over properties here must also include *complex* properties (e.g., '$\lambda x(P(x) \wedge \neg Q(x))$') for otherwise we cannot account for all the possible combinations of merely compossible properties that an object might have. For example, if we do *not* allow quantification over complex properties, we could still have models with only two worlds, one where every object has every property that is compossible with the causal core of the kind that it belongs to and one where it lacks all the compossible properties.

possible worlds.[16] That is, being a member of a kind, K, means having a causal core, C_K, and having this causal core in all causally possible worlds. As I assume that the metaphysically possible worlds are a superset of the causally possible worlds, this is compatible with metaphysical essentialism, which holds that it is metaphysically necessary for members of a kind to have their causal core (or essence).[17]

Hierarchy of Kinds

In order for premise (3) of the $\Diamond$-argument to be justified, I rely on a particular technical notion of kindhood: fundamental kind. To make clear what this notion is, we need to go over the idea that kinds (can) form a hierarchy (Tobin, 2010; Bird & Tobin, 2018). For example, all members of the kind KING PENGUIN are also members of the kind PENGUIN. To make things a bit more concrete, consider the representation in Figure 8.1. The causal core that is associated with the kind KING PENGUIN is the set of properties A, B, and D and the causal core of the kind PENGUIN consists of the properties A and B. It is easy to see that members of the kind KING PENGUIN are also members of the kind PENGUIN. So, a kind is *further down* the hierarchy of kinds if its set of causal core properties is a *superset* of that of its 'parent kind'.

For our purposes, I will define the notion of a *fundamental kind*, K_f. A fundamental kind is such that there is *no* further kind whose set of causal core properties are a superset of it. In our toy example, the kind PENGUIN is *not* a fundamental kind, as there are further kinds (e.g., KING PENGUIN, but also EMPEROR PENGUIN) whose causal core is a superset of the causal core of PENGUIN. KING PENGUIN, in

[16] It is not entirely clear whether, e.g., Khalidi (2013, 2018) would accept this kind of modal implication. On the other hand, Mallozzi (2018a) and Godman et al. (2020) accept something even stronger, namely that members of a kind have the causal core in all *metaphysically* possible worlds. Additionally, there is a worry, often raised by opponents of metaphysical essentialism, that there may not be necessary and sufficient conditions for being a member of a kind. They argue that there might not be a modally fixed set of properties that all members of, e.g., TIGER have. Relatedly, Khalidi (2013, 2018) points out that there might be kinds with fuzzy boundaries (2013, sec. 2.4). These issues raise the question of whether members of a kind should have the causal core by any way of necessity. For our purposes, I will assume that at least in all the causally possible worlds, members of a kind have the properties that make up the causal core. Fuzziness or flexibility (i.e., change) in the defining features of a kind can then be accommodated in causally impossible, metaphysically possible worlds. Another option would be to suggest that the issues of fuzziness and changing features of a kind give rise to *fallibilism*. In that case, strictly speaking, members of a kind needn't have the properties of the causal core in all the causally possible worlds, but they usually do. Thus, when reasoning about kinds, we assume that members of a kind have their core properties by way of causal necessity, but this reasoning is then fallible. These subtleties deserve close scrutiny in future work, but for the purposes of this dissertation I will set them aside.

[17] See Godman et al. (2020, sec. 7) for some arguments to the effect that members of a kind have their causal core in all metaphysically possible worlds. In Section 8.7 I will argue that this is too strong, but let me stress here that the arguments of this section justifying premise (3) are compatible with both views.

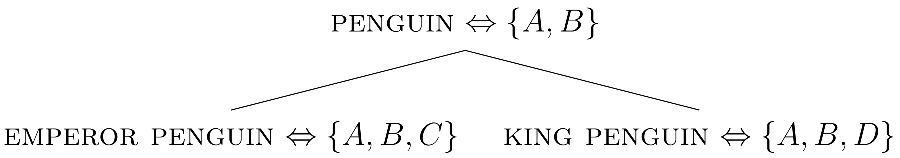

PENGUIN $\Leftrightarrow \{A, B\}$
EMPEROR PENGUIN $\Leftrightarrow \{A, B, C\}$
KING PENGUIN $\Leftrightarrow \{A, B, D\}$

Figure 8.1: *Toy example of a hierarchy of kinds.*

this example, is a fundamental kind as there are no further kinds that are 'below' it in the hierarchy.[18]

Finally, let me stress an important feature of fundamental kinds. Remember that members of a kind have their causal core by way of causal necessity. A fundamental kind is causally necessarily associated with *only* those properties that are in their causal core. That is, there are no other properties that a fundamental kind has in *all* the causally possible worlds. It is easy to see why this only holds for fundamental kinds and not for non-fundamental kinds, because for fundamental kinds there are no further kinds whose set of causal core properties are a superset of the causal core properties of the fundamental kind.[19] For even if, in our toy example, all members of the kind PENGUIN have their causal core in all causally possible worlds, some members will have an additional property, e.g., C, in all causally possible worlds (namely, members of the kind EMPEROR PENGUIN).[20]

8.2.2 Fundamental Kinds and Generalisations

So, being of a fundamental kind, K_f, means that it is causally necessary to have certain properties (i.e., the causal core) and that only those properties are causally necessary to have. Given all the above, we can now explain how this allows us to justify premise three: extrapolating from knowing that one fundamental kind member has a particular property to the fact that it is possible for each member of that fundamental kind to have that property. Consider the toy example of the penguins again. King penguins are members of a fundamental kind and properties A, B, and D make up their causal core. So, if we see that a particular king penguin

[18] Again, this feature of kinds is independent of the simple causal theory of kinds. All that is needed, is that kinds can form a hierarchy that can be understood in terms of set containment (see Bird & Tobin, 2018, §1.1.1). For example, metaphysical essentialists might hold that the essences of kinds are spelled out in such away that they also allow for the hierarchy of kinds.

[19] This says nothing about whether or not members of a kind belong to that kind necessarily$_M$. In particular, it leaves open the possibility that, e.g., Aristotle belongs to the kind HUMAN in this world, but to the kind DOG in another possible world (Mackie, 2009). I take it that these kinds of claims (e.g., 'Aristotle could be a dog') are outside the scope of modal knowledge we can get through the cognitive capacities that we use to acquire everyday, mundane modal knowledge (see Chapter 11).

[20] I set aside properties such as 'being-such-that-2+2=4' et cetera. Ultimately, we need a suitable notion of something like 'natural' properties.

has a further property, e.g., P, then we know that having property P is *not ruled out* by the set of properties $\{A, B, D\}$. Given that all king penguins only have these properties in their causal core and thus only have these properties by way of causal necessity, it is causally possible for king penguins to have property P.[21]

More generally, we say that, given the notion of a fundamental kind and its characteristics, if a member of a fundamental kind, K_f, has property P, then P is compatible with the causally necessary properties of being a member of K_f. As these are the only properties that are causally necessary for members of fundamental kind K_f to have, there will be no necessary properties that could defeat having property P. Therefore, it will be possible for all the members to have P.[22] There might be other properties such that having them would prevent one from having property P. However, by definition, members of fundamental kind K_f only have these properties contingently. So, given causal plenitude, there are worlds where they do not have these properties and do have property P. Thus, they *could* have property P.

Note that strictly speaking this inference only allows us to conclude that it is *causally possible* for king penguins to have property P. However, since (I assume) the causally possible worlds are a subset of all the metaphysically possible worlds, causal possibility implies metaphysical possibility.[23] (I will use the subscript '$_C$' for causal modality – e.g., 'it is possible$_C$ that' – and '$_M$' for metaphysical.)

We can make the case in favour of premise (3) of the $\Diamond$-argument more precise as follows:

(i) Being a member of a fundamental kind, K_f, means having certain (core) properties, $\{C_1, \ldots, C_n\}$ *necessarily$_C$* as well as those being the *only* properties that members of that fundamental kind have necessarily$_C$.

(ii) If a member of K_f has a (further) property P, then P is compatible with having properties $\{C_1, \ldots, C_n\}$.

[21] This reasoning doesn't follow from logic alone, in the sense that there could be models where there is only one world, in which case the reasoning doesn't go through. This is why we need causal plenitude, I will discuss this in more detail in Section 8.2.3.

[22] One might wonder what happens if the property in question is not just compatible with the causal core of the fundamental kind, but is actually *part of* this causal core. That is, what if $P = C_i$, where $C_i \in \{C_1, \ldots, C_n\}$. If this is the case, then it is in fact causally *necessary* that members of the fundamental kind have this property. However, remember that we are interested in the epistemology of *possibility* and even if objects have a particular property by causal necessity, it would still be causally *possible* that they have that property. (This holds for any kind of modality, χ, where the accessibility-relation is serial: $\Box_\chi \varphi \vDash \Diamond_\chi \varphi$.) It is irrelevant whether we *know* that the property in question is causally necessary for members of the fundamental kind, what we are interested in is that it is causally possible. This thus poses no problem for our account.

[23] This assumption would be false – i.e., $\Diamond_C \varphi \nvDash \Diamond_M \varphi$ – if there are worlds that are causally possible, but metaphysically *im*possible. I take it that this is highly implausible given that I take causal modality to be an objective modality and that metaphysical modality is the most objective modality (Williamson, 2016b) or modality *tout court* (Van Inwagen, 1998).

(iii) As all the members of K_f *only* have $\{C_1, \ldots, C_n\}$ necessarily$_C$ and P is compatible with those, all the members of K *could have (had)$_C$* P.

(iv) Since causal possibility implies metaphysical possibility, all members of K_f could have (had)$_M$ property P.

8.2.3 Some Initial Worries and Responses

The argument in favour of premise (3) is a valid argument based on what we take fundamental kinds to be; from this definition, it follows almost immediately. However, there are some initial worries that one might have concerning these notions and the resulting argument. I spell out the worries and dispel most of them here. Readers who think that the notion of a fundamental kind is not problematic and who want to focus on the justification for the premises of the $\Diamond$-argument can skip this section and move on to Section 8.3 on page 167.

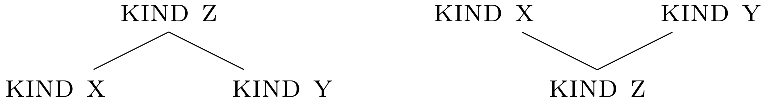
Modal Scarcity Worry

With the full argument for premise (3) on the table, we can now see why we need to assume causal plenitude. In particular, we can motivate it by looking at a potential worry for when we wouldn't assume it. I call this the *modal scarcity worry.* Consider the following scenario:[24]

> Theorist S holds that there is only *one* possible world: the actual world. That is, modal space is very scarce. So, any property that objects have, they have by way of (causal) necessity (at least their time-slices do). For any property P' and any object x, if x has P', then x has P' necessarily (in any sense of necessity, but we will focus on causal necessity).
>
> Now, consider two penguins, a and b – i.e., they are both of the kind PENGUIN. Further, a has the property of having albinism, which (let's assume) is neither a primary nor a secondary property of members of PENGUIN. It doesn't follow, according to S's metaphysics, that it is possible that b also has albinism. In fact, if b does *not* have albinism in the actual world, then it is causally necessary – on S's modal metaphysics – that b does *not* have albinism.

What this example shows is that the generalisation described by premise (3) does not follow from the metaphysics of kinds alone. We also need an additional metaphysical assumption about modal space, in particular about causal modality. What we need to rule out, as the modal scarcity worry shows, is the sceptical possibility of there being *no* other possible worlds (or too few possible worlds). This is why we assume

[24] Thanks to Peter Hawke for bringing this issue and the consequences it had for an earlier version of the argument to my attention. The example is the one raised by him.



(a) Hierarchical Kinds (b) Crosscutting Kinds

Figure 8.2: *Relation between kinds.*

the causal plenitude postulate: for each (merely) compossible property, each kind, and each member of that kind, there is a possible world where the member has the compossible property and one where it does not (see the precise formulation on page 159).

Crosscutting Kinds Worry

The definition of fundamental kind hinges on, what many have called, *the hierarchy thesis*: objects can belong to multiple kinds but only if these kinds form a nested hierarchy (Tobin, 2010; Bird & Tobin, 2018). In our example, a king penguin belongs to multiple kinds (e.g., KING PENGUIN and PENGUIN), but these kinds are nested. However, some people – for example, Khalidi himself – have *rejected* the hierarchy thesis and argued that there are *crosscutting kinds* – i.e., kinds that *partially* overlap (see Khalidi, 1993, 1998; Tobin, 2010; Khalidi, 2013). Figure 8.2 shows the difference between these two; with Figure 8.2(a) showing a hierarchical ordering of kinds, whereas Figure 8.2(b) is an example of crosscutting. For example, in Figure 8.2(a), members of kind Y also belong to kind Z, which, according to the hierarchy thesis, is fine as kind Y and Z form a nested hierarchy. However, in Figure 8.2(b), members of kind Z are also members of kinds X and Y, but kinds X and Y *do not* form a nested hierarchy (see Khalidi's real-life example discussed in the next paragraph). The question is whether the acceptance of crosscutting kinds affects the notion of fundamental kind in such a way that the above argument fails.[25]

However, crosscutting kinds do not seem to affect *fundamental* kinds. To speak within the metaphor of hierarchy, crosscutting kinds are such that a fundamental kind can be a subkind to multiple kinds; yet the fundamental kind in question always seems to remain a fundamental kind (according to my definition). To see this, consider an example from Khalidi: "the category *parasite* crosscuts the category *insect*. Tapeworms and fleas are classified together as parasites, and fleas and flies are both classified as insects, but neither category, *parasite* and *insect*, includes all three" (2013, p. 40, original emphases). The kind PARASITE and INSECT cut across each other as FLEA is a subkind of both, but they cannot form a hierarchy as they each include further subkinds that the other does not include (TAPEWORM

[25] Thanks to Tuomas Tahko for bringing this issue to my attention.

and FLY respectively). Though note that there are *no* further kinds that are more fundamental than, e.g., FLEA. That is, FLEA, in this toy example, is a fundamental kind however further classified, either hierarchically or crosscutting. That is, the notion of fundamental kind still is essentially the same as before and the argument holds.[26]

Necessary Defeaters Worry

Thirdly, one might wonder what happens if there is a *defeating* property, Q, that is incompatible with the projected property, P, and members of the fundamental kind have this defeating property necessarily by some *other* necessity than causal necessity (e.g., logical, metaphysical, nomic, etc.). If there is such a different necessity – i.e., such that $\Box_\chi \neq \Box_C$ – there are two ways in which it can relate to causal necessity – i.e., $\Box_C$. First of all, the necessity might be 'weaker' in the sense that if something is causally necessary it is also necessary by this other modality, but not *vice versa* – i.e., $\Box_C\varphi \vDash \Box_\chi\varphi$. That is, the χ-worlds are a proper subset of the C-worlds. In that case, even if all χ-worlds are such that fundamental kind members have a defeating property Q, there are still C-worlds where members of the fundamental kind do not have that property. Members of the fundamental kind $could_C$ still have property P that is incompatible with property Q, namely in a non-χ, C-world.

On the other hand, the necessity in question might be 'stronger' than causal necessity: if something is necessary according to this other modality, it is also causally necessary – i.e., $\Box_\chi\varphi \vDash \Box_C\varphi$. In this case, because it would also be causally necessary for members of a fundamental kind to have the defeating property Q, it would have to be part of the causal core of the fundamental kind. So members of the fundamental kind could not have property P in the actual world to begin with. Either way, this is not a problem for our theory.

Causal-to-Metaphysical Possibility Worry

Another worry one might have concerns the inference from causal possibility to metaphysical possibility. Let me just briefly mention two things in relation to this. First of all, there are many epistemologies of possibility in which a move very similar to this one is required (e.g., Hawke, 2011; Strohminger, 2015; Roca-Royes, 2017; Vetter, 2017). That is, it is not a problem for my particular theory; so, we can adopt the story concerning the justification of this inference given by these other theories. For example, Vetter (2017, sec. 3) suggests that we get knowledge of metaphysical modality from ordinary 'can' statements (e.g., 'I can reach for the door'), by extending the context of use for these statements. Alternatively, one might think that this inference is irrelevant for the kind of modal reasoning done by ordinary humans

[26] See Khalidi (2013, sec. 3.6), where Figures 3.1-3.4 are further examples that suggest that the crosscutting does not happen at the level of fundamental kinds, but only at the kinds that 'include' them.

and think that causal (or nomic) modality is the kind of modality that humans are (mostly) concerned with. Nothing of my argument hinges particularly on this choice: either the argument is done at step (iii) or we plug in a method of justifying the move from a natural or nomic possibility to metaphysical possibility.

Compatibility Worry

Step (ii) of the argument justifying premise (3) relies on a *compatibility judgement*. One might worry that explaining our judgements of possibility in terms of compatibility is too close of a circle.[27] The worry being that compatibility is closely related (perhaps too close) to notions that themselves involve modal judgements (for example, *consistency* as ‘not necessarily leading to a contradiction’, i.e., ‘being compossible’). However, it is not obvious that the compatibility judgements used here rely on (problematic) modal judgements. The relevant compatibility judgements are based on actuality – i.e., the argument for premise (3) gets off the ground if we know that there is a member of K_f that *actually* has property P. That is, we only need to know that the properties $\{C_1, \ldots, C_n\}$ and P are co-instantiated. This allows us to conclude, based on generalisation, that having properties $\{C_1, \ldots, C_n\}$ *does not rule out* the having of property P. If the worry is that this kind of judgement involves some sort of modal judgements, the worry seems to me very weak, as the modal judgement involved is very minimal (similar to the inference from actuality to possibility, see Chapter 1, Section 1.4.2).

Additionally, many suggest that compatibility judgements are part of our *basic cognitive machinery*; crucial for our survival in the (actual) world (see, e.g., Price, 1990; Berto, 2015; Berto & Restall, 2019).[28] For example, Price (1990, p. 226) argues that “sense of incompatibility” has a plausible evolutionary explanation, namely, it is required for successful signalling. “To signal significantly one needs to be capable of discrimination. One needs to signal in some circumstances and to remain silent in others. One needs a sense that these are mutually exclusive possibilities” (Price, 1990, p. 227). So, even if the relevant compatibility judgement is in some sense modal, it is such weak modal knowledge (compatibility in the actual world) that it does not undermine our argument. Our appreciation of compatibility and incompatibility is basic and primitive and we use it to judge what properties are compatible, in the actual world, with a cluster of properties exhibited by a kind.

Fundamental Kinds Worry

Finally, one might worry that our definition of being a member of a fundamental kind is not a very intuitive notion. This is true, it is a technical notion designed

[27] Thanks to John Divers for raising this worry.

[28] These authors argue for compatibility as the basis for a modal analysis of negation. Even those who disagree with such a modal analysis of negation, agree with the claim that (in)compatibility judgements are fundamental (De & Omori, 2018).

to make the argument work. This means that more needs to be said about how it relates to the notion of 'natural kind' and whether humans are able to know of or reason about fundamental kinds. These are legitimate questions and I will discuss these in detail in Section 8.4. For now we set these issues aside and move on to the categorisation premise of the $\Diamond$-argument.

8.3 Premise (1): Categorisation

The justification of premise (3), discussed in the previous section, was based on the *metaphysics* of kinds. The justification of the first premise of the $\Diamond$-argument will be based on the *epistemology* of kinds: categorisation. Premise (1),

$$\mathbf{1}.\ \exists K(Ka \wedge Kb)$$

requires us to categorise two things as belonging to the same kind. I will first discuss empirical literature that shows that we are generally very good at such classification tasks. Afterwards, I will argue that there are a number of ways in which this data can be used to justify the acceptance of premise (1); both for internalist and externalist epistemologists.

As with the hierarchy of kinds, there are multiple levels of classification. So, as we saw above, something that we might correctly classify as a KING PENGUIN, could, equally correctly, be classified as PENGUIN, BIRD, or ANIMAL. This raises a number of questions, two of which, for our purposes, stand out. First of all, is there a level of classification at which we are particularly good at classifying objects and, secondly, is there a particular level of classification at which we reliably *project* properties? The answer to both these questions is: yes. And, as it turns out, the level at which we are reliably good at classifying objects *is* the level at which we are the most reliable at performing inductive inferences. This may not be surprising as many have taken something like classification to be crucial for our ampliative reasoning (Anderson, 1990; Coley et al., 1999; Millikan, 2000; Gelman, 2003).

8.3.1 Categorisation and Classification

Let's start with our ability to classify and categorise the objects around us, before we turn to categorisation for the purposes of ampliative reasoning. Imagine walking into the Penguin Palace at your local zoo, encountering a large group of distinct objects (i.e., the penguins). "[I]f we responded to each object that we come across as if it were a unique individual, we would be overwhelmed by the complexity of our environment" (Markman, 1989, p. 11). So, in order to efficiently go through the Penguin Palace and, more generally, the world we *categorise*. "Categorization, then, is a means of simplifying the environment, of reducing the load on memory, and of helping us to store and retrieve information efficiently" (ibid.). Take another example, adapted from Millikan (2000). You are craving for some nice whisky and you happen to have a bottle of Bowmore in your pantry. If you are unable to categorise the content of that bottle as whisky, knowing that you have that bottle would not seem relevant for your whisky-craving. As Millikan puts it, both pieces of information are useless "unless you also grasp that these two bits of knowledge are about the same stuff" (2000, p. 6). Categorisation allows us to combine and relate distinct pieces of information that are about the same thing(s).

We start categorising the world around us from the very moment we enter it. Infants, for example, recognise that bottles are for feeding, whereas blankets are not (obviously, without linguistic attributions) (Markman, 1989). A bit later, we start categorising objects at the *basic level*. The basic level is the level of categorisation that is optimally informative; so categorising objects at a more specific level would only be insignificantly more informative and categorising them at a more general level significantly drops the informativeness of the categorisation (Rosch et al., 1976). For example, categorising something as bottle is very informative, whereas categorising it as whisky bottle is not much more informative and categorising it as object is very *un*informative. Children from about 18 months and older start categorising objects at this basic level (Markman, 1989, p. 15). We quickly start using these categorisations to interact with the world and throughout our lives, we categorise the objects around us based on increasingly specific and diverse categorisations. As Gelman and Meyer put it:

> Categorization takes place when an infant separates out carrots from peas on her dinner plate; when a toddler says 'doggie' in the presence of dog pictures, toy dogs, and the family pet; when a teenager decides which classmates are 'emos', 'jocks', or 'nerds'; and when a chemist identifies the elements in a sample of rock. (2011, p. 95)

Traditional Piagetian theories of human development suggested that at the early stages of our development we categorise things *purely* on perceptual similarities. However, during the second half of the twentieth century, this traditional view has been overthrown (see for example Rosch, 1978; Smith & Medin, 1981; Carey, 1985; Markman, 1989; Keil, 1989; Rips, 1989; and Smith & Sloman, 1994; for excellent contemporary reviews see Gelman, 2003 and Carey, 2009). Rips (1989), for example, provided very strong evidence that categorisation could not be *purely* based on similarity. Rips presented participants with a description (e.g., 'an object 8cm in diameter'), followed by two categories (e.g., pizza and quarter) such that one of the two categories is more variable than the other with respect to the property in the description (in this case, pizza is more variable in its size). Participants almost always picked the variable category to match the description. Rips hypothesised that this could not be explained by similarity, but *could* be explained by 'rule-based inferences' (e.g., 'a quarter cannot be more than 2cm in diameter'). Through these findings, Rips showed "a dissociation between categorization and similarity", which many take to be "the best documented dissociation of this sort" (Smith & Sloman, 1994, p. 378). These findings, though potentially the best documented, are one among many. In a recent overview article on categorisation, Gelman and Meyer note that,

> Adults' categories do not reduce to perceptual features alone; instead, they reflect domain-specific knowledge and theories [...]. Similarly, even 2-year-olds categorize objects based on functional features that conflict with surface appearances, [...]. Likewise, 3- and 4-year-olds categorize objects based on causal features, as long as the causal links are clearly and consistently demonstrated. (Gelman & Meyer, 2011, pp. 96-97)

However, Rips' wedge between categorisation and similarity should not be taken to show that similarity is *never* relevant for categorisation. For example, Smith & Sloman (1994) set out to replicate Rips' experiment with 'richer' descriptions (e.g., 'an object 8cm in diameter that is silver coloured'). In these cases, participants seemed to perform similarity-based, rather rule-based, reasoning. What Smith and Sloman conclude is that "this dissociation [between similarity and categorisation] occurs only under special circumstances. One such circumstance, or constraint, is that the description of the to-be-categorized object has to lack features that are characteristic of potential categories" (1994, p. 383). The debate between rule-based (or sometimes called 'theory-based') categorisation and similarity-based categorisation turns out to be extremely subtle and difficult to settle. Deák & Bauer (1996) therefore suggest that "instead of asking when and how preschoolers overcome perceptual boundedness, we [should] seek to *determine the circumstances* in which subjects of different ages use different kinds of information to make categorization decisions" (p. 742, emphasis added).

I can remain agnostic about all the subtleties surrounding our categorisation ability (is our skill to categorise universal? innate? what is the correct theory of it?) and merely focus on the fact that we are very good at categorisation. As some of the developmental data discussed above shows, we are categorising objects around us all the time in order to ease the cognitive burden of going around in this world. Moreover, there is widespread consensus that our ability to categorise objects is *fundamental* to our general cognition.[29]

- "[T]he child [...] is found to be a highly competent concept former" (Nelson, 1974, p. 272).
- "The results of the present experiment support the notion that humans are competent processors of [...] their knowledge about the degree of membership in the class of birds, furniture, and so on to perform their complex judgment task in a consistent and systematic fashion" (Oden, 1977, p. 201).

[29] See Deacon (1997) for evidence that this capacity is also present, to some extent, in primates (e.g., pp. 79-92).

- “As many investigators have pointed out, categorization is a fundamental cognitive process, involved in one way or another in almost any intellectual endeavor. [. . .] [T]he importance of categorization is very clear: most of human cognition depends on it” (Markman, 1989, p. 11).

- “[T]he cognizing organism must be able to recognize the specific substance under a variety of different conditions, as many as possible” (Millikan, 2000, p. 33).

- “Categorization is the mental operation by which the brain classifies objects and events. This operation is the basis for the construction of our knowledge of the world. It is the most basic phenomenon of cognition, and consequently the most fundamental problem of cognitive science” (Cohen & Lefebvre, 2005, p. 2).

- “But, at bottom, all of our categories consist in the ways we behave differently toward different kinds of ‘things,’ whether it be the ‘things’ we do or do not eat, mate with, or flee from, or the things that we describe, through our language, as prime numbers, affordances, absolute discriminables, or truths. And, isn’t that all that cognition is for – and about?” (Harnad, 2005, p. 40).

- “Categorization is a process that is intrinsically tied to nearly all aspects of cognition, and its study provides insight into cognitive development, broadly construed” (Gelman & Meyer, 2011, p. 95).

Categorisation might be one of the most fundamental human cognitive processes, essential for our survival. One of the main reasons why categorisation is so important, is its relation to *ampliative inferences*. In the next section, I will discuss this feature in more detail, as this will be crucial for the role of categorisation in the $\Diamond$-argument.

8.3.2 Categorisation and Induction

Consider the following example of the different levels of classification. When we come across a rainbow trout, we can classify it at, at least, the following levels:

Kingdom:	ANIMAL
Life form:	FISH
Generic:	TROUT
Specific:	RAINBOW TROUT

Given that we can classify a rainbow trout at all these different levels, it would be informative to know if there is any level that is particularly fruitful when it comes to ampliative inferences – i.e., the kind of inference we use in premise (3) of the ◇-argument. This is exactly what John Coley and colleagues studied: "[b]y examining the perceived strength of inductive inferences to categories of different folkbiological ranks, we hoped to discover whether one hierarchical level *is psychologically privileged with respect to induction*" (1999, p. 210, emphasis added). Based on prior research of, e.g., Rosch et al. (1976), who argued that there is a privileged level of classification that is optimally informative, Coley et al. hypothesised the following:[30]

> [I]nductive inferences to a privileged category should be significantly stronger than inferences to more general categories, but not significantly weaker than inferences to more specific categories. [...] In other words, we will consider the most specific level in folk biological taxonomy above which a significant breakpoint in inductive strength occurs to be inductively privileged. [...] [A] privileged level would be the highest or most abstract level at which inductive confidence is strong. (1999, p. 210)

In order to test their hypothesis, Coley and colleagues asked participants "to rate the relative strength of inferences from taxa of one rank to taxa of the next higher rank" (1999, p. 211). For example, in the rainbow trout example, they would be asked questions like: 'All trout have property P, how likely is it that all fish have property P?' and 'All rainbow trout have property Q, how likely is it that all trout have property Q?'. The participants were asked to rate the likelihood on a 9-point scale, 1 being 'not very likely' and 9 being 'extremely likely'. By comparing these answers, Coley and colleagues "were able to get a good look at the perceived strength of the inductive inferences" (1999, p. 211). The results are depicted in Figure 8.3 and

> clearly show that folk-generic categories (e.g., *trout*, *oak*) were inductively privileged for both the Itzaj and for American undergraduates. For both groups, inferences to folk-generic categories were consistently stronger than inferences to more general (life-form or folk-kingdom) categories, and no weaker than inferences to more specific (folk-specific) categories. (Coley et al., 1999, p. 211)

In combination with the fact that we are good at classifying objects, these findings of Coley and colleagues show that when it comes to classifying objects *for the purposes of ampliative reasoning*, humans focus on generic-level classifications (e.g., TROUT, PENGUIN, etc.). These findings are in line with other research on privileged levels of classification (see Rosch et al., 1976 for the privileged level with regards to informativeness; Keil, 1989; Medin, 1989; Coley et al., 1999; Gelman, 2003 for

[30] There are some subtle further motivations behind their research, namely that they were trying to bridge conflicting empirical results related to potential cultural influences on these kinds of inferences. These subtleties do not concern us here.

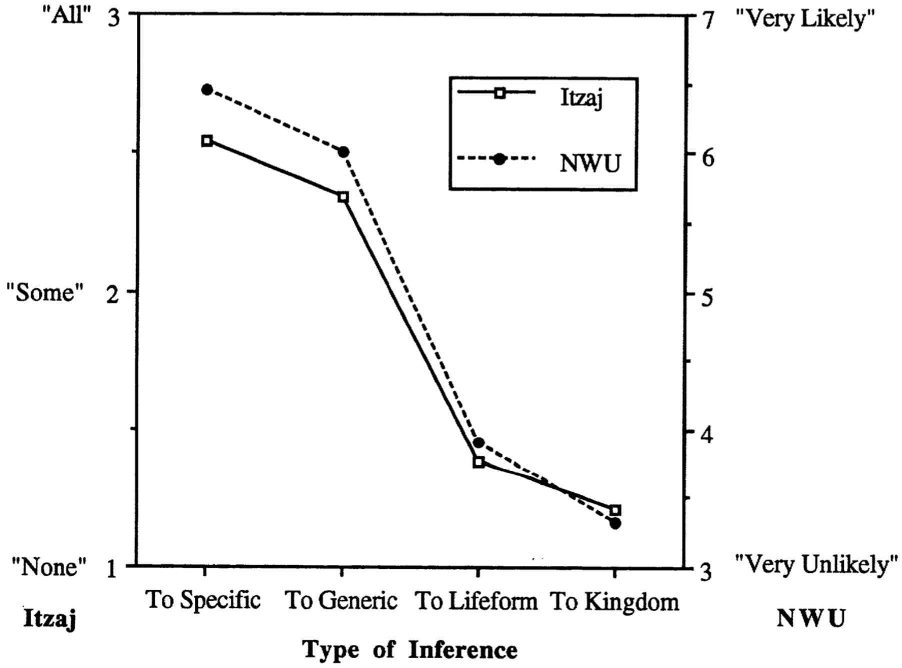


Figure 8.3: *Induction patterns from Northwestern Students and Itzaj Maya compared (from Coley et al., 1999, p. 212).*

findings related to privileged levels concerning ampliative inferences; and Prasada, 2000 for a suggested influence of language acquisition on this basic level).[31]

8.3.3 Epistemology of Categorisation

Given that the empirical data show that we are good at classifying objects, in particular at the generic level when we are concerned with ampliative inference, how can we use it to justify (the acceptance of) premise (1) of the ◇-argument? I will conclude the discussion of premise (1) by suggesting three possible ways in which one might take the relevant judgement – i.e., the 'sameness of kind' judgement – to be justified (or, how one might justifiably believe that two objects belong to the same kind). The first one is neutral on the internalism/externalism divide, but relies on a particular view concerning the content of perception (Siegel, 2016). The second and third are ways for the internalist and externalist, respectively, to use the data discussed above to justify premise (1). This discussion is not supposed to be exhaustive, in the sense that these are the only ways in which internalists and

[31] Note that it is quite obvious that this generic level of classification does not (neatly) match the fundamental kind level used in premise (3) (for one thing, one would expect RAINBOW TROUT to be a fundamental kind in this example). I will elaborately discuss this apparent mismatch in Section 8.4.

externalists can appeal to the data. Rather it is meant to show that there is at least one way in which they could justify premise (1) given the empirical findings – a proof of concept if you will.

Categorisation through Perception

The content of our perceptual experiences contains, at the very least, qualities such as colour, shape, distribution in space, et cetera. However, some have argued that beyond this, the content of our perceptual experiences also include (i) representations of ordinary objects and (ii) kind properties (e.g., Siegel, 2006, 2016).[32] The representation of basic objects in perceptual experience, though a bit more controversial than the representation of the primary qualities, is still widely accepted. If it is indeed true that we see basic objects, and we assume that we are justified on the basis of perception, we are justified in believing (the content of) premise (1) for a large range of kinds. As Siegel notes, "[t]he class of ordinary objects is notoriously difficult to define, but it is clear enough to support theorizing by psychologists [. . .]. And plenty of examples of ordinary objects can be given—*cats, keys, tables, and the like*" (2006, p. 482, emphasis added). We would be justified to believe premise (1) on the basis of perception for even more objects if it is true that we see *sortal* properties (Siegel, 2006 calls this 'Thesis K'). In that case, "visual experience [also] represents properties such as *being a house, and being a tree*" (Siegel, 2006, p. 483, emphasis added). As Siegel (2016, §4.3) points out, it is controversial whether or not these sortal properties are part of the content of our perceptual experience.

Examples that are often given in favour of Thesis K concern cases "in which the subject's beliefs about what she is seeing seem to affect visual phenomenology" (Siegel, 2006, p. 489). In order to explain this, consider a famous example of the *opposite effect*: the Müller-Lyer Illusion, where even after one is told that the lines are of the same length, they continue to look as if they are of different lengths. The subjects' beliefs *do not* affect their visual experience. The arguments in favour of Thesis K concern cases where subjects' beliefs *do* seem to affect their experiences. Consider the following example (adapted from Siegel, 2006, p. 491):

> You are hired by your local zoo to take care of *only* the Emperor Penguins. The zoo has both Emperor and King penguins, which look very much alike, and you have no prior experience with penguins. So, at first, you fail to see the difference between Emperor and King penguins and your colleagues have to help you, by pointing out the Emperor penguins. After a couple of months, you are sent to another zoo, to help them with the care of their Emperor penguins. Even though this zoo also has both Emperor and King penguins, you now have no difficulty to see which

[32] Some have even argued that we see modal properties such as objects being 'edible' 'climbable', etc. (see Nanay, 2011a,b; Strohminger, 2015).

> penguins are Emperor penguins and which ones aren't.[33]

In Siegel's terms, "your disposition to distinguish the [Emperor penguins] from others improves. Eventually, you can spot the [Emperor penguins] immediately. They become visually salient to you" (2006, p. 491). Your visual experience seems to be affected by your "recognitional disposition" (ibid.); the sortal property of 'being an Emperor penguin' seems to be part of your visual experience (see Siegel, 2006, p. 491ff. for a full defence of Thesis K based on such examples).

If we experience basic objects and sortal properties, then we get justification for something *stronger* than premise (1) directly from perception. We would be able to directly categorise objects that we perceive, which is stronger than merely being able to judge two objects to be of the same kind (which allows one to remain ignorant of *which* kind they are).

Yet, as already noted, including representations of basic objects and sortal properties into the content of perceptual experience is controversial. So, how might one justify the acceptance of premise (1) if one *rejects* this? I will discuss how internalists and externalists might do so in turn.

Categorisation, Internalism, and Game-Theoretic Warrant

One method that internalists might use in order to justify the sameness judgement is by appeal to, what Wright (2004, 2014) calls, *entitlement of rational deliberation*.[34] Such entitlement is a form of *non-evidential warrant* according to Wright. These are "grounds, or reasons, to accept a proposition that consist neither in the possession of evidence for its truth, nor in the occurrence of any kind of cognitive achievement [...]. Still, a non-evidential warrant is warrant to accept a proposition as true" (Wright, 2014, p. 214).

Wright's entitlement theory is, roughly, a form of *hinge epistemology* (see Wright, 2014, sec. 11.1 & 11.2). He phrases his approach in terms of cognitive projects ("defined as a pair: a question, and a procedure one might competently execute in order to answer it" Wright, 2014, p. 215) and *authenticity-conditions* for such a cognitive project. The authenticity-conditions are such that doubting them undermines the entire project.

> [A]n authenticity-condition for a given cognitive project is any condition doubt about which would rationally require doubt about the efficacy of

[33] The addition of you being transferred to a different zoo, is to block the possibility that you recognise the Emperor penguins because you have become familiar with the particular, individual penguins at your own zoo and not in virtue of you seeing that they are Emperor penguins.

[34] Roca-Royes, when talking about the justification of ampliative reasoning, notes that her "working hypothesis is that a sceptical solution in terms of *entitlement of rational deliberation* is, not just the best we can do, but all we need" (2017, p. 232, original emphasis).

> the proposed method of executing the project, or about the significance of its result, irrespective of what that result might be.
>
> (Wright, 2014, p. 215)

The method for identifying which propositions deserve warrant (justification) based on entitlement through rational deliberation can differ. Sometimes, it is suggested that there is a sort of *indispensability* of the cognitive project or authenticity-conditions (Philie, 2009); other times it is argued that it concerns proposition III of a I-II-III scepticism (Wright, 2004; Philie, 2009; Wright, 2014);[35] and, finally, it has been suggested that it concerns *dominant strategies* in a Reichenbachian, game-theoretic approach (Wright, 2014).

An internalist could appeal to any of these methods to argue for the applicability of entitlement through rational deliberation for the sameness judgement in premise (1) (or, in the case of the indispensability approach, the justification for the stronger 'correct categorisation'). I will discuss two of these. In general, one way to think of it is that the proposition 'These two things are of the same kind' is a cornerstone (hinge) proposition for induction and cognition in general.

A kind of indispensability argument to justify the sameness judgement would, just like the perception approach discussed above, proceed by justifying the stronger claim, namely that categorisations are, generally, justified. The reasoning would go as follows. Rational cognition as a whole is a cognitive project for which categorisation is an authenticity-condition.[36] Given that rational cognition is indispensable and categorisation a necessary condition for cognition (as we saw in the previous two subsections), categorisation is justified. This would be the kind of indispensability entitlement approach that one could make.

Secondly, one could appeal to Wright's Reichenbachian approach to entitlement of rational deliberation, which focuses directly on the sameness judgement, rather than on justified categorisations (see Wright, 2014, sec. 11.3 for a detailed explication of this and references to Reichenbach's work). On such an approach, one has

[35] I-II-III scepticism, for example in the case of induction, goes as follows:

I All observed F's are G's

II All F's are G's

III Nature is uniform

Where proposition I justifies proposition II and proposition II justifies proposition III. However, the sceptic will point out that in order for I to justify II, one already has to accept proposition III. This results in a circularity that undermines not only propositions of kind III, but even mundane propositions of kind II (e.g., in the Moorean case, proposition II would be 'I have hands'). Wright (2004, 2014) argues that we might be 'warranted' to accept propositions of kind III based on entitlement of rational deliberation.

[36] Alternatively, one could take any cognitive process that relies on sameness judgements (e.g., ampliative reasoning, recognising food, etc.) as the relevant cognitive project.

	Nature is uniform	Nature is haphazard
Trust in truth-conduciveness of induction	**Many true and useful beliefs**	Few true and useful beliefs
Lack of trust	Few true and useful beliefs (or many true and useful beliefs, but irrationally due to lack of trust)	Few true and useful beliefs

Figure 8.4: *Wright's (2014, p. 227) game-theoretic matrix for induction.*

to argue that *trust* in a particular cognitive project (or authenticity-condition) is a game-theoretic *dominant strategy*. That is "[i]n all relevant possible futures, the mooted course of action either works out better than all alternatives or no worse than any alternative" (Wright, 2014, p. 224). If it is the case that trust in a cognitive project is the dominant strategy, then we are justified in trusting that particular cognitive project and thus justifiably accepting its 'outputs'. For example, Wright suggests that we can use the game-theoretic matrix depicted in Figure 8.4 to justify our use of induction (where the dominant strategy is in boldface). Similarly, we can create a game-theoretic matrix in order to justify the trust in our capacities of judging things to be of the same kind. One such matrix is shown in Figure 8.5, though it is likely that there are other matrices to the same effect.

	Are of same Kind	Are not of same kind
Trust in sameness judgement	**Many true and useful beliefs**	Few true and useful beliefs
Lack of trust	Few true and useful beliefs (or many true and useful beliefs, but irrationally due to lack of trust)	Few true and useful beliefs

Figure 8.5: *Game-theoretic matrix for sameness judgement.*

Not all of these entitlement approaches to the justification of the sameness judgement of premise (1) might be equally appealing. Nor might it be that an entitlement approach in general is the best way for an internalist to justify the sameness judgement. However, this preliminary discussion does show that the internalist has at least one promising option to develop an epistemology of categorisation that results in the justification of the acceptance of premise (1).

Categorisation, Externalism, and Reliabilism

Externalists, on the other hand, can rely on a variety of closely related approaches. I will briefly discuss *reliabilism* (but see Goldman (2012, ch. 3) for an overview of similar externalist accounts). On such an approach, an agent is justified in believing something if that belief was caused or formed by a process that is reliable, i.e., that in general produces true beliefs (see Goldman, 1979).[37]

One of the motivations for developing reliabilism is a dissatisfaction with some aspects of internalism. For example, on an internalist account, agents are supposed to be able to 'access' whatever it is that justifies their beliefs (see, e.g., BonJour, 2003; Pappas, 2017). So, if I believe φ on the basis of some perceptual experience, I should be, according to internalists, be aware of the fact that it is that particular perceptual experience that justifies my belief in φ. According to reliabilists, and externalists in general, this is too strong a requirement. Consider young children or some animals: we would want to say that they hold certain justified beliefs (e.g., about their favourite toys), but it seems unlikely that they know *what it is* that justifies them having that belief. Additionally, Goldman (1979) discusses a number of, roughly internalist, attempts to define when a belief is justified and argues that most of these definitions are *circular* in that they rely on epistemic terms in the definition. Goldman suggests that what goes wrong is "that each of the foregoing attempts confers the status of 'justified' on a belief without restriction on why the belief is held. [. . .] I suggest that the absence of causal requirements accounts for the failure of the foregoing principles" (1979, pp. 8-9).

In response, externalists deny that one has to be aware of or to be able to access that which justifies their beliefs. For example, reliabilists suggest that as long as the belief arises "from the deployment of mental processes and methods, [. . .], that are conducive to acquiring true belief and avoiding error in actual and/or modally relevant circumstances," then the belief is justified (Goldman, 2012, p. 3). The agent in question does not need to be aware that these reliable processes justify their beliefs. In particular, Goldman proposes a *process* reliabilism,[38] where beliefs are justified if the process that produces them (in our case, the process of categorisation) is reliable (Goldman, 1979, p. 13).[39]

[37] Fricker suggests that reliabilism is "the dominant theory in contemporary analytic epistemology" (2016, p. 88).

[38] "Let us mean by a 'process' a *functional operation* or procedure, i.e., something that generates a *mapping* from certain states – 'inputs' – into other states - 'outputs'. The outputs in the present case are states of believing this or that proposition at a given moment" (Goldman, 1979, p. 11, original emphasis).

[39] Later, Goldman (1992, ch. 9) went on to propose *two-stage* reliabilism, where first the ascriber

Given that we very reliably classify objects in order to use them in ampliative inferences, it seems clear that on a reliabilist account, we are justified in accepting (the content of) premise (1).[40]

8.4 The Placeholder Heuristic

We have a justification for premise (3) of the $\Diamond$-argument (i.e., valid reasoning from the definition of what a fundamental kind is) and we have evidence that we reliably make correct classification judgements, which we can use to justify premise (1). However, it is not obvious that the technical notion of a *fundamental kind* is something that humans use in reasoning. This suggests that the empirical evidence on categorisation at the generic level might be *irrelevant* for premise (3) of the $\Diamond$-argument. To see this, let us dub the level in the hierarchy of the metaphysical kinds that corresponds to the generic level at which we reliably make correct classifications the *basic level* (or *basic kind*). Then, if we use subscript variables for the kinds, K, in the relevant ways, the $\Diamond$-argument looks something like this:

1. $\exists K_b(K_b a \wedge K_b b)$
2. Pb
3. $\forall K_f \forall P'(\exists y(K_f y \wedge P'y) \rightarrow \forall x(K_f x \rightarrow \Diamond P'x))$
Con. $\Diamond Pa$

It is no longer obvious that the argument is valid and a worry arises concerning a potential equivocation between the fundamental kinds in premise (3) and the basic kinds in premise (1). Making the reasonable assumption that the generic level at which we make categorisations is in fact *not* the level of fundamental kinds (as the example of the rainbow trout suggests), there seem to be two initial responses to the equivocation in the $\Diamond$-argument.

First of all, one might 'weaken' premise (3) so that it concerns basic kinds rather than the fundamental kinds:

$$\forall K_b \forall P'(\exists y(K_b y \wedge P'y) \rightarrow \forall x(K_b x \rightarrow \Diamond P'x))$$

This move is problematic, as the premise is now no longer true. To see this, consider the following example. Let us assume that MAN and WOMAN are two fundamental kinds 'under' the basic kind HUMAN. Thus, the set of causal core properties of WOMAN is a superset of the causal core properties of HUMAN (the same goes for the set of causal core properties of MAN). Now consider the property of having two

composes a list of (un)reliable processes and, secondly, checks if the belief-formation process was reliable. This development is meant to deal with evil demon- and clairvoyance-cases (see also Fricker, 2016).

[40] Interestingly, this is similar to reliabilist approaches to the problem of induction that deny the viciousness of *rule-circularity* (Van Cleve, 1984; Papineau, 1992).

X allosomes (i.e., sex chromosomes), which is a necessary$_C$ property for being a member of WOMAN, but not for being a member of HUMAN (for men don't have this property). So, if we generalise on the level of basic kinds (in this case, HUMAN) and we know that there are some members of HUMAN that have two X allosomes (e.g., we know some women have these properties), we would, falsely, conclude that all members of HUMAN *could have* two X allosomes. Reasoning according to this weakened premise (3) would thus make predictions – e.g., that men could have two X allosomes – that we should not accept. At best, this might be true but not something we should want our theory to predict to be knowable (as Roca-Royes, 2017 concluded with the different parents example). At worst, it is actually impossible for men to have two X allosomes, in which case our theory would predict something false. Avoiding this is the reason why we restrict premise (3) to fundamental kinds.

Secondly, we might suggest to 'strengthen' premise (1), so that the categorisation concerns fundamental kinds rather than basic kinds (or generic level categorisation):

$$\exists K_f(K_f a \wedge K_f b)$$

The problem with this option is that it is not obvious that we make reliable judgements about fundamental level classifications. For example, the data from Coley et al. (1999) show that people focus on projections at the generic level, rather than at the specific level, and one might rightly wonder whether or not it actually is the specific – rather than the generic – level that corresponds to the fundamental kinds (or perhaps there is an even more specific level). Similar studies on categorisation (e.g. Rosch et al., 1976; Gelman, 2003) suggest that humans reason at a level of categorisation that is *not* the most specific (i.e., fundamental) level of classification. To refer back to our toy example from above, just think how good the average human is in categorising penguins (i.e., members of the basic kind PENGUIN) versus how poorly most humans would be in categorising emperor and king penguins (i.e., respectively, members of the fundamental kinds EMPEROR PENGUIN and KING PENGUIN).

I suggest a different approach. One that is, again, inspired by empirical evidence concerning the way humans reason about and with kinds (Medin & Ortony, 1989; Strevens, 2000; Gelman, 2003; Cimpian & Salomon, 2014). What this evidence, which we will discuss in more detail shortly, shows is that humans often reason *as if* the level at which they categorise objects is the level at which objects have '*essences*'. This is known as *psychological essentialism*, a psychological theory about the way humans reason about kinds. This is clearly distinct from and not to be confused with the *metaphysical* essentialism about kinds (discussed in Appendix C). In order to keep things clear, I will write that humans reason as if kinds have 'essences', letting the scare quotes indicate that humans are *not* concerned with the (neo-)Aristotelian essences of metaphysical essentialism.

Given that humans reason as if the level at which they reliably categorise objects is the level at which objects have an 'essence', we can take humans to be reasoning as if their level of categorisation is the level of fundamental kinds. The best way to cash this out, I suggest, is by explicitly incorporating a *heuristic* into the rational reconstruction.[41] This revised rational reconstruction – let us call it the $\Diamond_H$-argument – looks as follows (again with the relevant subscripts to indicate the level of the kind):

1. $\exists K_b(K_b a \wedge K_b b)$
2. Pb
3. $\forall K_f \forall P'(\exists y(K_f y \wedge P'y) \rightarrow \forall x(K_f x \rightarrow \Diamond P'x))$
4. $K_f b \approx K_b b$
Con. $\Diamond Pa$

The crucial, new, assumption is premise (4). What this is meant to represent is the, what I will call, *placeholder heuristic*. In ampliative reasoning based on kinds, humans reason *as if* the basic kind, at which the categorise the objects, allows for unrestricted projection of properties; humans reason as if the basic kind is the fundamental kind.[42] As this is a heuristic and not actual identity, I use '$\approx$' to indicate this relation (I've phrased things in terms of object b, given that the projection is based on the object of which we know it has the property in question, but the same holds for object a after having classified it as of kind K).

Adding the placeholder heuristic allows us to leave premises (1) and (3) as they were, involving basic and fundamental kinds respectively, and so we can keep the justification for these premises as before.

In this section, I will first spell out what I take the placeholder heuristic to be, based on the *causal placeholder theory*. I will argue, based on empirical data, that humans do reason according to the placeholder heuristic and that there are good, evolutionary reasons for doing so. I will conclude by highlighting the heuristic aspect of it: it is a shortcut of getting at shared deep, 'essential' properties and this shortcut gets things right most of the time. This last part is discussed in more detail in the next section, where I turn to the types of mistakes that result from this defeasible heuristic reasoning.

8.4.1 Causal Placeholder Theory

There are a number of different theories that hold something along the following lines: people reason as if there is a crucial, core cluster of properties that 'make'

[41] I mean to follow the standard usage of 'heuristics' as, for example, in discussions concerning bounded rationality, where they take heuristics to be "simple rules of thumb for rendering a judgement or making a decision" (Wheeler, 2020, §7).

[42] Note that there is another way of capturing this heuristic reasoning: all the reasoning actually happens at the basic level of classification and we add an assumption concerning the generalisations and projections at this level. I hypothesise that this would result in roughly a similar account, however, fully working this out is left for future work.

members of a kind belong to that kind (e.g., Medin & Ortony, 1989; Prasada, 2000; Strevens, 2000; Gelman, 2003; Cimpian & Salomon, 2014). For example, there is something that makes cats be members of the kind CAT. Varieties of *psychological essentialism* claim that humans reason *as if* there is an 'essence' that kinds have: it is in having a particular 'essence' that cats are members of the kind CAT (Medin, 1989; Medin & Ortony, 1989; Gelman, 2003).[43] According to others, this is too strong, we don't ascribe 'essences' to kinds, we merely reason as if kinds have a common causal core (Prasada, 2000; Strevens, 2000; Cimpian & Salomon, 2014). They suggest that we can explain all the data that psychological essentialists appeal to *without* ascribing judgements of 'essences' to people; all we need to explain the data is that humans ascribe a core set of causal properties to members of a kind.[44]

What is crucial is that all these theories hold that we *don't* need to know exactly what this 'essence' or common causal core is, we merely believe that it is there (Prasada, 2000, p. 67). That is, all these theories hold that we have a *placeholder* for whatever the exact core cluster of properties is – i.e., people reason as if members of a kind have a core, "even if its details have not yet been revealed" (Gelman, 2003, p. 10). In the previous chapter (in Section 7.3), I argued that reasoning by predictive analogies results in justified beliefs if we know the *explicit causal relations*, i.e. knowledge of all the particular causal relations involved. In the case of these psychological theories, the required prior knowledge is weaker: we do not need to know *exactly* what the relevant causal properties are. We just know (or reason as if we know) that there is some causal core that consists of some properties, even though we may be unaware of which properties exactly and which relations hold between them.

For our purposes, we don't need to accept psychological essentialism and we can just accept the minimal claim: humans reason as if there is a core cluster of properties common to all members of a kind that allows for (unrestricted) ampliative reasoning. Let us call this *the Causal Placeholder Theory* (CPT for short). This is closely related to a number of theories in the field, which. I will briefly discuss. My argument is compatible with any of these theories; if it turns out that one of the other theories is true, we can simply adopt that one.

CPT is very similar to Strevens' (2000, p. 154) *minimal theory*: we "believe there are causal laws connecting natural kinds and their observable properties, but [we] are not committed to any particular view about the implementation of these laws." However, the reliance on the common core consisting of causal *laws* is more specific than we need for our purposes.[45] Similarly, my suggestion is closely related

[43] The kind of essentialism that these theorists ascribe to is *psychological, causal, placeholder* essentialism (see Gelman, 2003, pp. 7-11).

[44] Note how nicely this psychological story aligns with the, metaphysical, Simple Causal Theory of kinds discussed above (Keil, 1995 also points out that a causal understanding of kinds nicely fits with these psychological theories).

[45] Khalidi (2018, p. 1389, fn. 9) notes that instead of characterising the relations between the

to the view of Cimpian, who holds that "there exists at least one feature that is typically shared" such that we use that feature "for the purposes of generating an in-the-moment explanation." These explanations, in turn, "tend to be about the *inherent* (constitute, stable) features of the entities" (2015, p. 5, original emphasis). Cimpian relies heavily on a particular view of our cognitive architecture (the dual-systems approach, see Cimpian, 2015, p. 4) that we need not accept. Finally, I take it that psychological essentialism accepts this minimalist characterisation and *additionally* holds that this cluster of properties represents an essence (see Strevens, 2000 and Cimpian & Salomon, 2014 for arguments to this effect). So, if all we need for our purposes is the CPT, then we can remain agnostic as to whether or not the surplus of psychological essentialism is true.

Triad Tasks

There is widespread consensus that something like CPT is true (see Gelman & Markman, 1986, 1987; Medin & Ortony, 1989; Medin, 1989; Gelman & Coley, 1991; Strevens, 2000; Cimpian & Salomon, 2014 and Gelman, 2003, ch. 2 for an excellent review). In order to get a feel for the kind of data in favour of it, let me briefly review the seminal *triad tasks* of Gelman & Markman (1986, 1987) (the presentation here is an abstraction of the actual empirical set ups).

The triad task is based on work by Carey (1985) on inductive inferences as a method for testing the nature and development of children's concepts. Gelman & Markman (1986, 1987) adapted Carey's experimental set-up in order to test the relation between induction and category membership. Findings from the triad tasks, relevant for the CPT, have been replicated many times and, in general, it has been found that "the effect is very robust" (Gelman, 2003, p. 31).[46]

The test itself can be conducted in a number of different ways, but always involves a *triad* of objects: a source object and two objects for comparison. One of the comparison objects would be perceptually similar yet of a different kind and the other would be perceptually dissimilar but of the same kind as the source object. For example, if the target object is a (black)bird, then the perceptually similar yet

core and derivative properties as causal, one might take them to be natural laws (Hawley & Bird, 2011). If one takes this approach, Strevens' approach might be favourable.

[46] Moreover, Gelman and colleagues were very careful with their methodology and explicitly worked to avoid the demand characteristic (see, e.g., Orne, 1962; Rosnow, 2002).

> My collaborators and I have also conducted numerous control experiments to rule out the possibility of task demands and to determine the role of category membership (as opposed to superficial matching strategies) on children's performance. One primary concern with the initial studies was that children may have made use of the category information simply because they were attempting to please the experimenter.
>
> (Gelman, 2003, p. 35)

This significantly solidifies their findings.

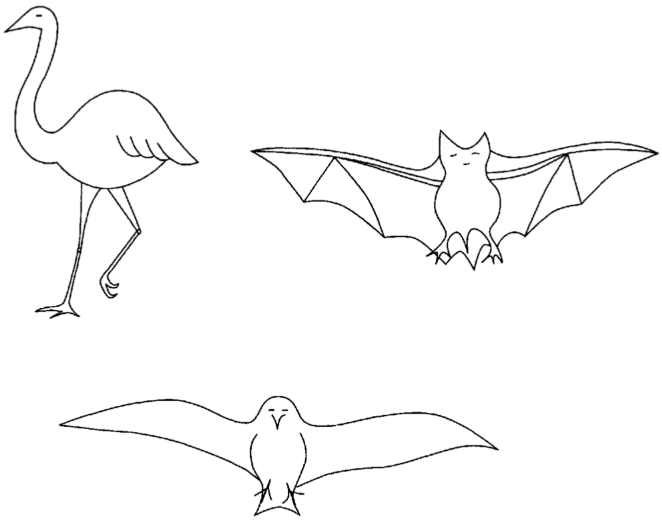


Figure 8.6: *Sample item set for triad task (from Gelman, 2003, p. 29).*

categorically distinct object would be a bat and the, perceptually dissimilar, category member would be an ostrich (see Figure 8.6; Table 1 of Gelman & Markman, 1987, p. 1536 and Table 2.1 of Gelman, 2003, p. 30 provide more examples of triads).

The experimenters make sure that the participants knew the correct categorisation (either by labelling the objects (e.g., for preschoolers) or by doing a control study). Then, and this is where instances of this experiment differ, the participants were submitted to one of two reasoning tasks. Either, they were told an unknown, novel property about the source object and then asked which of the two comparison objects would also have that property. For example, they would be told that the (black)bird has property P and then asked whether the bat or the ostrich has property P. Otherwise, they would be told two distinct properties of the two comparison objects and then asked which of these two would most likely apply to the source object. So,

> the children were told, ‘This fish [i.e., tropical fish] stays underwater to breathe’ and ‘This dolphin pops above the water to breathe.’ After being shown the picture of the shark and told that it was a fish, the participants were asked whether it stayed underwater to breathe like the fish or popped above the water to breathe like the dolphin.
>
> (Gelman, 2003, p. 29)

Tests like these have been done on participants ranging from very early ages of development (e.g., two years Gelman & Coley, 1990; three years Gelman & Markman, 1987; four years Gelman & Markman, 1986) to undergraduates and adults (Gelman, 2003). In all cases the results are clear, people favour category membership over perceptual similarity in reasoning inductively (Gelman, 2003, p. 30).

What is important for us, in support of CPT, is the *reasons* that participants gave for their choices. Remember, CPT suggests that humans reason as if belonging to a kind implies having a particular set of structural (causal) properties that allows for ampliative inference. So, the kind of justification that would be in favour of CPT would be that participants point to something intrinsic that is particular to members of that kind. Gelman (2003) notes that adults do just this; they would justify their inferences with explanations "such as 'Birds are *structured internally* alike,' 'Usually animals of the same species have similar characteristics,' and 'Gold is gold' " (Gelman, 2003, p. 30, emphasis added).[47]

The triad tasks are perhaps the most clear cut evidence that humans reason according to CPT (for a full review of the empirical evidence see Gelman, 2003, ch. 2, especially pp. 27-43). Overall, the combined data suggests that something like CPT holds for both adult and children's ampliative reasoning. "The only explanation that satisfactorily accounts for the varied patterns of data is that children assess the extent to which entities are members of the same kind" (Gelman, 2003, p. 59).

8.4.2 The Placeholder Heuristic

So, participants often appealed to, something like, CPT *in order to justify* their ampliative inference. But, as we saw above (at the beginning of this section), generalising to all members of a kind at the basic level allows for counterexamples. The fundamental level of kinds, as we saw in Section 8.2, does not; it is valid to extrapolate to all its members. One explanation of the confidence with which people extrapolate to other members of a kind is that people reason as if this causal core they ascribe to the kind does not allow for defeaters. That is, they reason as if the causal core at which they categorise an object is the core of a fundamental kind. (For more evidence on people's confidence on ampliative reasoning about kinds, see Coley et al., 1999 and Gelman, 2003, ch. 2.)

Why think that we should accept reasoning based on heuristics and not just dismiss it as faulty reasoning?[48] I suggest that this reasoning is correct most of the time, in particular, enough of the time for it to be evolutionarily beneficial for us (for *this* heuristic, I do not want to commit to a view on reasoning based on heuristics in general). I will discuss this motivation for accepting the reliance on the

[47] Gelman & Coley (1991, pp. 168-169) provide more evidence of the kind of justification that people give for their reasoning in triad tasks that is in line with CPT (especially their Table 5.1).

[48] Compare the main disagreement between the two schools on heuristics and biases, which concerns this normative question (Stein, 1996; Vranas, 2000; Wheeler, 2020). Whereas the school following the work of Tversky & Kahneman (1974, 1983) (see also Kahneman, 2011) describes the heuristics they discuss in terms of 'errors' and 'fallacies', those following the work of Gigerenzer (1996, 1998) (see also Gigerenzer & Murrya, 1987) think they are characteristics of our adaptive human psychology (see Stein, 1996 and Vranas, 2000 for excellent reviews of this normative discussion).

placeholder heuristic in the remainder of this section and focus on the mistakes due to this heuristic in the next section.

Gelman (2019) notes, "heuristics don't have to be 100% right—they just have to be right enough of the time to allow for fruitful predictions. Our reasoning heuristics can lead to errors, yet do the work to get us to survive another generation and even be a boost to learning" (p. 327). When it comes to the ampliative reasoning based on our categorisation into kinds, this is exactly what people have suggested: we gained this skill through evolution as a way to move safely through the world (Quine, 1969, p. 14; Medin, 1989, p. 1477; Kornblith, 1993, p. 104; Millikan, 2000, p. 146; and Gelman, 2019, p. 316). In particular, many have suggested that this type of reasoning has evolved *as a heuristic* (Medin, 1989; Cimpian & Salomon, 2014; Cimpian, 2015; Gelman, 2019). People use CPT-like reasoning to make quick judgements of sameness, of classification, and – most importantly for us – to make inductive inferences.

When reasoning according to a heuristic is justified, is something that is almost never discussed. I will not attempt to solve the justification of heuristics here, but let me just hint at some initial thoughts. In line with our externalist justification for premise (1) (i.e., the sameness judgement), we might appeal to something like process reliabilism. The data from this section shows that reasoning according to the causal placeholder heuristic gets things right most of the time, or at least enough of the time to be evolutionarily useful. A reliabilist might suggest that that is enough to justify relying on such a heuristic in reasoning and the beliefs that follow from such reasoning are justified. For internalists, it is less obvious if they could justify the reliance on heuristics.[49] It seems to me that, as in the case of the sameness judgement, internalists might appeal to something like rational entitlement through the game-theoretic evaluations of trust in heuristics. I will assume that the data discussed above are enough to at least make it plausible that the reasoning results in justified beliefs. A full epistemology of heuristics will be left to another occasion.

8.5 Fallibilism: What mistakes can we make?

We now have the full picture of the theory that I propose, the $\Diamond_H$-argument. When we try to determine whether it is possible for an object, a, to have a particular property, we reason based on kind judgements. We consider whether we know of any objects that are of the same kind and that have the property in question. If so, we conclude that a *could* also have the property. In reasoning like this, we rely on,

[49] It is of course a potential point of debate for everyone whether we *want* to say that the beliefs one acquires based on heuristics *are* justified. I leave this discussion for future work (for a somewhat related debate, see the normative debate in the field of bounded rationality; see references in footnote 48).

what I have called, the placeholder heuristic: we reason as if the kinds we reliably recognise allow for generalisation to all its members.

By relying on a heuristic, the resulting reasoning is *fallible*: even though the $\Diamond_H$-argument justifies our beliefs in possibility claims, it does not guarantee their truth, i.e., we might be wrong (see the discussion of fallibilism in Chapter 1, Section 1.3.3, see also Leite, 2010; Brown, 2018). In this section, I will discuss the particular kinds of mistakes that we might make while reasoning according to the $\Diamond_H$-argument. The placeholder heuristic allows for two kinds of mistakes. Either we *overestimate* the range of potential generalisation – i.e., it turns out that the generic level at which we categorised the kind members is not the level at which we can unrestrictedly generalise. Or we *underestimate* the range of potential generalisation – that is, we take certain properties to belong to the causal core of a kind when they do not.[50] I will discuss these two kinds of cases in reversed order.

8.5.1 Underestimation and Epistemological Modesty

The reason why we start with the *under*estimation of the range of projection, is because this is the unproblematic one. Consider our example of the penguins again (Figure 8.1 above and Figure 8.7 below) and imagine a situation where we have seen members of multiple different specific kinds of penguin, but we've only seen king penguins swim (for some reason, all other penguins were seen on land). Based on this we conclude that objects that share the causal core of KING PENGUIN can swim, but *not* that all members of PENGUIN can swim. In doing so, we *underestimate* the range of this generalisation. As it turns out, the property 'can-swim' is compatible with the causal core not of any *fundamental* kind, but with the basic kind, PENGUIN; all penguins can swim.

In cases of underestimation, our reasoning is not really defeated (nor really fallible), for it is not *wrong*. We are simply being *epistemically modest*. We do not infer more than the evidence allows us to, even though we would not be wrong if we did. In particular, my proposed epistemology of possibility turns out to be *modally modest* (in the sense of Hawke, 2011, 2017): we can know mundane, everyday possibility statements when we are acquainted with 'relevantly similar' instances, but we are *not* justified in believing exotic possibility statements, far removed from ordinary experiences (e.g., the existence of philosophical zombies). Note that this kind of modesty is similar to the cases mentioned in the literature. For example, Roca-Royes (2017) suggests that if we only have seen IKEA-tables break, we should conclude that IKEA-tables can break, but not that tables in general can break, even

[50] Note that my terminology here differs from Gelman (2019), whose data I report here. She labels the first kind of mistake an 'underestimation of the variability' and the latter kind an 'overestimation of the importance of category boundaries'. I prefer my labels as they relate the mistakes more clearly to the heuristic of taking basic levels to be fundamental kinds.

though 'breaking' might be compatible with the causal core of tables in general.[51] Similarly, Hawke (2017), also dealing with inductive inferences, suggests that we should *not* conclude that emeralds can be yellow based *only* on other, non-emerald, diamonds that are yellow (even though 'yellow' might be compatible with the causal core of diamonds in general).

There is ample empirical evidence of this kind of underestimation; mostly coming from cases of sexism and racism. People tend to exaggerate the differences between race and sex, even if there are actually no such differences (see Gelman, 2019, sec. 12.5). That is, these people *underestimate* the range of their projection. For example, in suggesting that *only* girls are fragile or like pink, they project these properties only to girls, whereas boys might equally be fragile or like pink. There is nothing in the causal core that prevents the projection of 'likes-pink' to members of BOY as well as of GIRL (Gelman, 2019, p. 318).[52]

Over time and with a growing understanding of the *content* of the causal core, rather than a blanket placeholder, we become more epistemically confident to make the projections at the right level.[53] By gaining knowledge of the *explicit* causal relations involved in the causal core of a kind and the property we are projecting, we slowly move towards reasoning by predictive analogy (see the previous chapter). In these cases we will become better and more confident in assessing at which level in the hierarchy of kinds we can generalise the property in question.

8.5.2 Overestimation and the Case of Defeat

The other type of errors we make concern *overestimation*: we mistakenly assume that our projections at the basic level (of categorisation) reflect the stability of fundamental kind reasoning. Consider the following real-life example of this type of mistake. In November 2000, a large number of penguins washed up on the beaches of Rio de Janeiro. Based on their understanding of penguins (basic level classification), people that were trying to help reasoned that these animals lived in environments with temperatures below freezing. So, they reasoned, putting these washed up animals in their freezers would save them. However, these particular animals – Magellan penguins – live in environments where the temperature *never* falls below freezing, resulting in many of the 'rescued' penguins being on the verge of dying.

> The Brazilians apparently assumed (falsely, as it turned out) that knowing that a bird is a penguin allows you to infer that its habitat and

[51] I haven't said anything about the distinction between natural and artifactual kinds. However, I assume for the purposes of this chapter that much of what I've said here also applies to artifactual kinds.

[52] There are some other issues that Gelman (2019) discusses. For example, the fact that humans seem to focus on dangerous features (as well as inborn or inherent) (p. 321). Further study would be needed to test how much these biases affect our projection of *modal* properties.

[53] Potentially helping us overcome some of these sexist and racist prejudices.

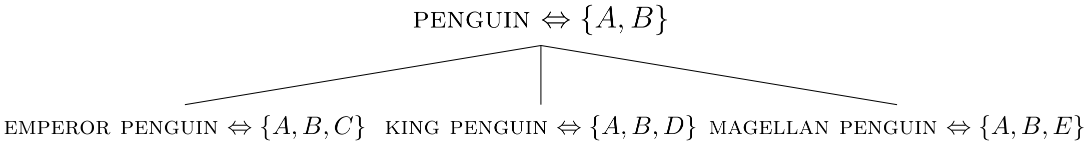

PENGUIN ⇔ $\{A, B\}$
EMPEROR PENGUIN ⇔ $\{A, B, C\}$
KING PENGUIN ⇔ $\{A, B, D\}$
MAGELLAN PENGUIN ⇔ $\{A, B, E\}$

Figure 8.7: *Toy example of a penguin hierarchy.*

> body temperature are equivalent to those of other penguins. They relied on what they already knew about a subset of the category to make inferences about novel category members. Unfortunately, the category of penguins does not cohere as tightly as the Brazilians' naive theories led them to believe. (Gelman, 2003, p. 26)

In order to pinpoint where things went wrong, it will help to picture things in terms of levels of classification and the causal core that the kinds corresponding to these levels have (as the toy example represented in Figure 8.7). The emperor and king penguins are probably the most famous and most paradigmatic kinds of penguins and they both live in environments with temperatures almost constantly below freezing. The magellan penguins, on the other hand, live a couple of hundred kilometres more towards the Equator, in environments where temperatures almost never fall below freezing. It is likely that, on the basis of their familiarity with emperor or king penguins, the Brazilians thought that all penguins need freezing temperatures to survive.

Their reasoning mistake was that they assumed that members of the basic kind PENGUIN – with properties A and B as their causal core properties – require environments with freezing temperatures, whereas actually, only penguins with the causal core properties $\{A, B, C\}$ and $\{A, B, D\}$ do so. Phrased differently, their generic level categorisation (i.e., PENGUIN) did not correspond to a fundamental kind (e.g., MAGELLAN PENGUIN) and thus did not allow for infallible projection of properties (see Gelman, 2019, sec. 12.3 & 12.4 for further examples).

Suitable Defeaters

Errors of overestimation are truly cases where the $\Diamond_H$-argument reasoning can be defeated, showing that the reconstructed reasoning is fallible. Given that we rely on our ordinary, everyday cognitive capacities, which are themselves fallible, this was to be expected (see also Williamson, 2007). An interesting question with regards to these overestimation errors is: what are suitable *defeaters* to the $\Diamond_H$-argument? The empirical findings concerning these reasoning errors focus on *the actual world*. As Gelman points out, her findings are evidence of the fact that reasoning according to CPT "oversimplifies the complexities of *the natural world*" (2019, p. 316, emphasis added). However, what we are after with the $\Diamond_H$-argument is something much weaker: *modal* knowledge concerning possibilities. We are not interested whether the magellan penguins in the actual world survive in temperatures below freezing, but whether they *could* do so. This gives us some hints as to what counts as a defeater for our purpose.

When one *over*estimates the range of the projection, there is a property in the causal core of a member of the fundamental kind that defeats the projected property. In our penguin toy example, having the additional property E, defeats the projected property of being able to live in temperatures below freezing. What the previous paragraph shows is that it should be *impossible* for the defeating property in question to be co-instantiated with the projected property. So, in our toy case, having property E makes it *impossible* to live in temperatures below freezing. Strictly speaking, these defeaters should make it *metaphysically* impossible that the projected property is instantiated. This means that the chances of defeat are even *less* than when we would focus on *causal impossibility* (i.e., there are worlds where the co-instantiation in question would be causally impossible, yet metaphysically possible). Given that we focus on everyday modal knowledge, plus the fact that our $\Diamond_H$-argument focuses on the causal aspects of kinds (both in the metaphysics and epistemology), it makes sense to accept properties that make it *causally* impossible to have the projected property as defeaters for the $\Diamond_H$-argument.

There is some evidence that we *do* reason about causal impossibility in such away (for example as discussed by Byrne, 2005; Nichols, 2006a; Rafetseder et al., 2010; Phillips & Knobe, 2018; Redshaw et al., 2018; Phillips et al., 2019; Leahy & Carey, 2020). For example, Shtulman & Phillips (2018) found that children often regard violations of physical regularities to be impossible. Relatedly, adults seem to make similar judgements when they are under time-pressure (Phillips & Cushman, 2017). An example of such an impossibility judgement is discussed by Nichols (2006a, p. 243). He suggests that when a child tries to determine whether it is possible for a hornet to get in a car and we "close off by stipulation" all the ways that a hornet might causally get in the car (e.g., not by magic), then the child will no longer conclude that it is possible for the hornet to get in the car. In general, these findings suggest that there is "a developmental pattern such that young children regard events involving physical violations [. . .] to be impossible" (Phillips & Knobe, 2018, p. 14). (Phillips and Knobe also present an excellent philosophical discussion of the earlier findings of Phillips and colleagues.) These findings support the idea that (judged) *causal* impossibilities defeat our ordinary possibility judgements.

Let me stress that these findings do *not* imply that we infer metaphysical *im*possibility from causal impossibility; nor do I intend to suggest this. Whether or not these causal impossibilities are also metaphysically impossible is not something that our theory should come down on.

There seem to be two main types of errors we can make in reasoning according to the $\Diamond_H$-argument: (i) we are modally modest, only projecting properties to members of a category even if it might apply to the members of the 'higher' level, and (ii) we sometimes overgeneralise, projecting properties to all members of a basic category, while the property only applies to members of some more specific categories. In the latter case, our reasoning is defeated if the properties these members have are causally (not just actually) incompatible with the projected property.

Should these mistakes worry us about the reasonableness of relying on something like the placeholder heuristic (as represented by premise (4) in the $\Diamond_H$-argument)? No. Following Gelman, who critically evaluated human reasoning based on such heuristic and pointed out many of the errors discussed above, we should conclude that despite these mistakes reasoning based on the placeholder heuristic is reliable enough for learning and evolutionary purposes (Gelman, 2019, p. 327).

8.6 Kinds and Possibility: Theoretical Virtues

We now have the full picture of the theory that I propose, as summarised at the beginning of the previous section. In this section, I will briefly consider the theoretical virtues of the proposed theory, in line with the guiding assumptions (e.g., a roughly naturalist, cognitively plausible approach) discussed in Chapters 1.

Fit with pre-theory

First of all, this account nicely aligns with the pre-theoretic description of what goes on in similarity-based reasoning. Once we know that certain objects of a kind have certain properties, then we know that there is nothing in being a member of that kind that prevents those objects from having the property in question. Consider the following dialogue about justifying that a coffee cup could break:

(a) This cup can break. **(b)** How do you know this?

(a) This is a particular kind of cup and these cups can break.

(b) How do you know these cups can break?

(a) I saw another cup of the same kind that did in fact break, thus there is nothing in these kinds of cups that prevents them from breaking.

The theory presented here nicely captures this kind of reasoning. (The same goes for the theories of Hawke, 2011 and Roca-Royes, 2017, where the dialogue would have to be phrased in more general terms of ‘relevant similarity’.)

Plausibly Cogent Reasoning

The conclusion of the $\Diamond_H$-argument follows from its premises. Moreover, the kind of reasoning involved in the $\Diamond_H$-argument (e.g., sameness judgements, ampliative inference based on the placeholder heuristic) are part of our ordinary cognitive machinery (hence cognitively plausible, see below). An example from Millikan (2000, p. 6) nicely captures how fundamental this kind of reasoning is:

> Suppose, for example, that you are hungry and that you know that yogurt is good to eat and that there is yogurt in the refrigerator. This is of no use unless you also grasp that these two bits of knowledge are about the same stuff, yogurt.

In our everyday lives, we rely on these kinds of judgements constantly. So, scepticism about the reasoning involved in the $\Diamond$-argument results in widespread scepticism about our ordinary cognitive facilities (see also Williamson, 2007, ch. 5). This means that sceptics about the reasoning in the $\Diamond_H$-argument would either have to bite the bullet and argue that we *in general* are not justified on the basis of these kinds of judgements or they would have to suggest that there is something special about this 'modal context' that makes the, normally reliable, kind of reasoning defective. Neither option seems attractive (Machery, 2017).

However, the reasoning is predicted to be defeasible; there is no infallible 'modal vision' or 'rationalistic intuition' or anything of that sort involved (e.g. Bealer, 2000). This leaves room for one to be sceptical about *far-fetched cases* or lack of conceptual competence/requisite background knowledge. That is, this leaves room for *modal modesty*, in the spirit of Van Inwagen (1998) and Hawke (2011) (see also Chapter 11 and Strohminger & Yli-Vakkuri, 2018b).

Cognitively Plausibility

Reasoning based on kind judgements as captured by the $\Diamond_H$-argument is taken seriously as a correct model of ampliative reasoning in the cognitive sciences and lends itself to an evolutionary explanation (as many of the references and quotes above testify). This is precisely the kind of methodologically naturalist, cognitively plausible approaches to philosophy that Williamson (2007) and Machery (2017) argue for. Nolan argues forcefully for the usefulness of naturalistic approaches to the epistemology of modality in particular, noting that even non-naturalists "stand to benefit from the further development of naturalistic methods in our modal investigations" (2017, p. 26). The account discussed above can be seen as complying with these suggestions for non-exceptionalism in philosophy and a cognitively plausible epistemology of modality.

Substantial/Objective Subject Matter:

The knowledge resulting from the $\Diamond_H$-argument, is knowledge of *objective* modality (Williamson, 2007, 2016b; Machery, 2017; Vetter, 2017; Strohminger & Yli-Vakkuri, 2018a). That is, modality itself is at issue, not just the concept of modality; metaphysical not conceptual modalities are in play. This is explicitly stressed by those working in the psychology of modality (e.g., Nichols, 2006b; Phillips & Cushman, 2017; Shtulman & Phillips, 2018; Phillips & Knobe, 2018; Redshaw et al., 2018). For example, Nichols notes that the empirical data he discusses explicitly concerns *objective*, rather than epistemic modality. “It’s noteworthy,” he points out, “that the children’s modal claims probably can’t typically be given a deflationary kind of epistemic interpretation. [. . .] [C]hildren seem to *deploy non-epistemic* modal notions” (2006a, p. 242, emphasis added). (See also Khalidi, 2013, ch. 2 on the relation between the epistemology of projection and the objective nature of kinds.)

8.7 Mallozzi and the Mismatch Worry

Let me conclude this chapter by discussing an epistemology of modality that is, perhaps surprisingly, rather similar to the one proposed here. Mallozzi (2018a) takes a similar starting point to that of this chapter, in that she focuses on our modal knowledge based on kinds and she also takes kinds to be based on a (simple) causal common core. What is interesting is that, where we focused on an epistemology of possibility, she focuses on an epistemology of *necessity* and relies heavily on metaphysical essentialism. First, I will discuss Mallozzi’s general approach and the epistemology of modality that she proposes. I will then note an interesting epistemological difference between our theories. Finally, I will raise a worry for her account that, I will argue, does *not* arise for my account.

Mallozzi advocates a *modal-metaphysics first* approach to the epistemology of modality. That is, she thinks that we should first know what metaphysical modality *is* before we can develop theories of how we come to know things about it. According to Mallozzi, metaphysical modality is grounded in essences: whatever something’s essence is accounts for what is metaphysically necessary for that thing (see Mallozzi, 2018b, pp. 7 & 15). According to her, this view allows for a blissful marriage between a Finean (1994) metaphysics of essences – i.e., the constitutive view of essences – and the Kripkean (1980) epistemology of necessities – i.e., relying on a bridge principle. All this cumulates in, what she calls, “the basic bridge-principle” that relates essences to necessities and, *a fortiori*, the knowledge thereof:

(E) “If it is essential to x being F that it is G, then necessarily anything that is F is G.” (Mallozzi, 2018a, p. 6)

So, from the epistemological side of things, when we know the essence of something, we (can come to) know the metaphysical necessities concerning that thing. For Mallozzi this means that we should “recast the epistemology of necessity in terms of the epistemology of essence” and focus on how we come to know the essence of things (2018a, p. 8).

It is at this point that the simple causal theory of kinds comes in. According to Mallozzi, the causal core of kinds (based on the work of, e.g., Craver, 2009;

Khalidi, 2013), simply *is* the essence of members of a kind. “The crucial point for our purposes,” she suggests, “is that this causally and explanatory powerful core is what I call the ‘essence’ of the kind, and thus what constitutes the fundamental nature of the kind” (Mallozzi, 2018a, p. 9).[54] Given that this causal core grounds the essence of a kind and that essences ground metaphysical necessities, Mallozzi holds “that many metaphysical necessities can be understood by applying this causal-explanatory notion of essence” (2018a, p. 12; see also Godman et al., 2020, p. 11).

The proposal in this chapter shares with Mallozzi (2018a) the reliance on a causal theory of kinds and we both think that modal knowledge is related to knowledge of kinds (for Mallozzi this is mostly relevant for modal knowledge of natural kind identities, e.g., ‘water is H_2O’ and ‘Gold has atomic number 79’). However, our theories are also significantly different. For one, Mallozzi’s view is a *Rationalist Two-Factor* view (Strohminger & Yli-Vakkuri, 2017), she relies on a bridge-principle and she suggests that our knowledge of such bridge-principles is due to “an *a priori* step of some sort (inferential or intuition-based)” (2018a, p. 17). My account, on the other hand, is firmly *empiricist*. Secondly, Mallozzi proposes an epistemology of *necessity*, whereas I suggest an epistemology of *possibility*. It may seem surprising that two views that both start from similar assumptions, still end up at such significantly different epistemologies of modality.

Let me first mention an interesting observation about the difference between Mallozzi’s epistemology of kinds and mine. For Mallozzi, we get to know the common causal core of a kind by “scientific investigation aimed at disclosing the causal structure of kinds” (2018a, p. 18). What she seems to have in mind is that we need to know the exact, explicit causal relations and properties to determine essences, which we can only find out through scientific investigation. This is different from the proposal in this chapter. First of all, nowhere in the $\Diamond_H$-argument does an agent need to know *what* kind the objects under investigation are; all they need to know is that the two objects are of the same kind. Moreover, where Mallozzi focuses on scientific investigation into the causal core of kinds, I focus on the reasoning on ordinary humans.[55] I suggest, based on empirical data from developmental and cognitive psychology, that even though humans reason as if there is a common causal core for kinds, they do not (need to) know the exact causal relations involved.

[54] In Godman et al. (2020), Mallozzi notes a subtle difference between her view and that of, e.g., Craver (2009) and Khalidi (2013). She notes that in her own view, “the great preponderance of natural kinds owe their clustering of properties, not just to some causal structure or other, but to *one single* underlying property that serves as the common cause of all the other clustered properties” (Godman et al., 2020, p. 6, emphasis added). Whether or not the causal core of a kind consists of a single property or a cluster of properties does not matter for my main objection against Mallozzi.

[55] Maybe, what Mallozzi has in mind for the justification of modal judgements of ordinary humans is something like an *epistemic division of labour* (to paraphrase Putnam, 1973).

The difference is not one of kind, but of degree. I agree with Mallozzi that scientific investigation will help us get a firm grasp on the causal core of kinds. However, in order to explain the modal knowledge that we have and use in everyday life, and that young children seem to have, we cannot expect this to be due to fully worked out theories of what these causal cores are. Instead, we rely on the placeholder heuristic in order to make reliable judgements without knowing the exact properties that make up the causal core of a kind. I take it that this sort of kind-based similarity relation lies on a *continuum* of similarity relations, where one of the extremes is the predictive analogy (as discussed in Chapter 7), which can be interpreted as judgements based on full (scientific) *understanding* (Gentner, 1983; Gentner & Markman, 1997).[56] However, I think that if we aim to explain our ordinary, everyday possibility judgements, requiring prior knowledge of the exact causal relations involved in the causal core is too high an epistemic demand.

8.7.1 The Mismatch Worry

If scientific understanding and the sort of kind-based similarity reasoning that I propose lie on a continuum, then you might be tempted to think that this type of reasoning can provide us with a *uniform* epistemology of modality: we use the proposal of this chapter to get knowledge of possibility through kind-based reasoning and we use Mallozzi's proposal to get knowledge of necessities based on the same kind of reasoning. Though tempting, I think that Mallozzi's suggestion that the *causal* core of kinds are *metaphysically* necessary is an unfounded mismatch between two kinds of modality. Let's phrase the *mismatch worry* as follows: why think that the common causal cores of kinds 'constitutively determine metaphysical necessities'?

Many researchers working on the philosophy and metaphysics of (natural) kinds, especially those working on a causal theory of kinds, *explicitly reject* that belonging to a kind is metaphysically necessary.[57] The main reason for rejecting these modal implications is that "it is not clear what grounds there are for holding that there is a general connection between a kind being natural and its applying to its members necessarily" (Khalidi, 2013, p. 28). The worry is that taking causal properties to be metaphysically necessary is unwarranted and that there is a mismatch between the

[56] Another very interesting epistemology of modality that relies on something like understanding, yet is significantly different from Mallozzi's, is that of Fischer (2016b, 2017b). He provides an analogy between the epistemology of modality and games: in order to know what possibilities there are for the game of *Clue* (or any other game), the most definitive way is to understand the rules of the game. From this, one can deduce whether or not something is possible in the game, namely if there are no rules that it violates. The same goes, Fischer argues, for *interesting* modal claims: we are justified in believing them if we believe a theory according to which these claims are true and we believe them on the basis of the theory (Fischer, 2017b, p. 8). I leave a full comparison with and evaluation of Fischer's work for future work.

[57] Godman et al. (2020) do admit that "[p]hilosophers of science who work on kinds [. . .] are generally resistant to any suggestion that they have 'essences' " (p. 2).

necessity Mallozzi takes essences to be related to – i.e., metaphysical – and the kind of necessity that causal cores of kinds might be related to – e.g., causal, natural, or nomic (see Priest, 2018 for similar worries).

In response, one might accept a *deflationary* account of metaphysical modality where metaphysical necessity *just is* the causal (or nomic) necessity that aligns with the necessity by which kinds have the properties of theirs causal core. Mallozzi responds to a similar objection (Godman et al., 2020, sec. 8), but it is not quite clear what this response comes down to. She admits that essences, on her account, have a nomological flavour and that she is trying to "reduce [metaphysical necessity] to a specific kind of nomological structure" (2020, p. 14). Yet, on the other hand, she keeps stressing that the causal core of a kind is a property that "will be possessed in all *metaphysically* possible worlds" (ibid., emphasis added). Whether or not Mallozzi would accept it, I take it that on such a deflationary account of metaphysical modality, it is no longer the case that metaphysical necessity is "distinct from other notions of necessity" in any interesting way (Priest, 2018, p. 1). If this is the case, then the epistemology of modality that Mallozzi (2018a) proposes does not get us knowledge of *distinctly metaphysical* necessities – i.e., metaphysical necessities that are nomologically impossible.

So, we can accept that knowledge of essences grounds knowledge of metaphysical necessities and that the causal core of kinds is necessary in some sense. Then we can either accept the deflationary account of metaphysical necessity discussed above, in which case the knowledge is not of distinctly metaphysical necessities, or we reject the deflationary account, in which case we need additional arguments to accept that there is a link between this causal necessity and metaphysical necessity. In this last case, without further argument there remains a mismatch between what she takes the essences of kinds to be and the kind of necessity that these essences are supposed to ground. The common causal core of kinds (potentially) ground causal necessities, yet Mallozzi needs her essences to ground *metaphysical* necessities.

One might wonder whether my account falls victim to a similar mismatch worry. However, given that I focus on *possibility* rather than necessity, I can accept that the causal core of kinds is related to nomic (or in my case, causal) modality rather than to metaphysical modality. Because nomic possibility *implies* metaphysical possibility, there is no problem here (e.g., Williamson, 2016b). This entailment, however, does not (straightforwardly) hold between nomic necessity and metaphysical necessity.

Potentially, metaphysical essentialists might worry that by rejecting the link to metaphysical necessity, we lose some of the epistemic power of our kind-based similarity reasoning. This is not the case. Rejecting this essentialist link to metaphysical necessity does nothing to undermine the epistemic fertility of kinds. For "there is widespread agreement among essentialist *and nonessentialist* philosophers alike that natural kinds are the grounds for rich inductive inferences" (Khalidi, 2013, pp. 14-15, emphasis added).

Part III

Philosophical Possibilities

Chapter 9

Gettier Reasoning and the Problem of Defeat

> *Much of the philosophical community allows that a judicious act of the imagination can refute a previously well-supported theory*
>
> – Williamson, 2007

In the previous two parts, we focused on a cognitively plausible epistemology of possibility that accounts for our justified beliefs in ordinary possibilities. In this chapter we will see that possibility judgements are also crucially important to the practice of (Western) analytic philosophy. In particular, they are a central part of one of philosophy's main tools: *thought experiments*. Recently, there has been an interest in analyses of the epistemology of these philosophical thought experiments that rely on a cognitively plausible epistemology. As mentioned in Chapter 1 (Section 1.3), non-exceptionalism is the idea that *philosophical thinking* does not require any cognitive capacities beyond those that we, humans, already possess for our ordinary, everyday interaction with the world (Williamson, 2007, p. 136).

In this chapter, I will discuss one of the main theories of the epistemology of thought experiments (Williamson, 2007), as well as the main objection to it: the problem of deviant defeat. I will argue that the objection applies much more widespread than is thought and actually affects *most* accounts of the epistemology of thought experiments. This insight allows us to focus on *other* important aspects of Williamson's account. The resulting discussion highlights an often overlooked question that is essential to the epistemology of thought experiments: how do we know that a hypothetical situation is possible? The positive proposals of epistemologies of possibility in this dissertation (in particular Chapters 5 and 8) strengthens the case for non-exceptionalist analyses of thought experiments.

9.1 Thought Experiments in Philosophy

Roughly, thought experiments are hypothetical situations about which one makes a judgement.[1] Some of the, perhaps, most famous thought experiments come from the sciences: Galileo's falling-bodies; Newton's climbing buckets; Maxwell's Demon; Einstein's elevator (and train); Schrödinger's cat; et cetera. Even though there is a lot of interesting work done on in the philosophy of *scientific* thought experiments, I will exclusively focus on *philosophical* thought experiments, which are thought experiments "put forward by philosophers" and "are almost always meant to elicit a judgment" (Machery, 2017, p. 11).

The use of thought experiments in philosophy is pervasive throughout the history of philosophy.[2] Cumulating in an impressive list of famous thought experiments in contemporary philosophy: Russell's (1948) clock; Wittgenstein's (1953) beetle; Quine's (1960) *gavagai*; Gettier's (1963) cases; Foot's (1967) trolley; Dretske's (1970) zebras; Thomson's (1971) violinist; Putnam's (1973) Twin-Earth; Goldman's (1976) fake barns; Kripke's (1980) Gödel; Searle's (1980) Chinese room; Jackson's (1986) Mary; Davidson's (1987) Swampman; Chalmers' (1996) zombies; and many more. As Stuart et al. (2018b, p. 1) put it in their introduction to *The Routledge Companion to Thought Experiments*, "[t]hese thought experiments in large part define the history of philosophy and are woven into its pedagogy." The use of thought experiments in philosophy is not surprising, Sorensen argues, for "it is the natural test for the clarificatory practices constituting conceptual analysis: definition, question delegation, drawing distinctions, crafting adequacy conditions, teasing out entailments, advancing possibility proofs, mapping inference patterns" (1992b, p. 15). In general, it is assumed that thought experiments "constitute a fundamental item within the bag of tools of analytic philosophers" (Angelucci & Arcangeli, 2019, p. 763). "Philosophy without thought experiments seems almost hopeless" (Brown & Fehige, 2019, p. 1).[3]

These thought experiments are used in a number of different ways. Some are merely used to raise questions, whereas others are used in support of a particular theory (Cohnitz & Häggqvist, 2018). I limit my discussion to the class of thought experiments that are used to *refute* particular (claims of) theories. For example, to take the case study of this chapter and the next, Gettier cases were meant to *refute* the claim that knowledge is justified true belief. Gettier did so by presenting hypothetical cases that (usually) invoke the judgement that the agent in the situation has a

[1] Some accounts of thought experiments start from a description that is more liberal (Machery, 2017; Cohnitz & Häggqvist, 2018) and others include more restrictions on the basic definition (Brown & Fehige, 2019).

[2] See Stuart et al. (2018a, especially essays in Part 1); Stuart et al. (2018b); and Brown & Fehige (2019, §3.1) for historical overviews.

[3] However, as we will see in Chapter 10, there is a dissenting voice coming from (a group of) experimental philosophers (e.g., Machery, 2017).

justified true belief, but we wouldn't judge the agent to have knowledge. This then is taken to show that there are cases where knowledge is not justified true belief and thus constitute a counterexample to any theory that claims this.[4]

There have been many labels for this class of thought experiments. For example, Popper (1959) called these *critical* thought experiments; Brown (1986) calls them *destructive*; Sorensen (1992b) uses the term *alethic refuters*; and Cohnitz & Häggqvist (2018) and Stuart et al. (2018b) call them *thought experiments as counterexamples*. I will avoid these labels and just talk of 'thought experiments,' even though in what follows I will only focus on thought experiments as counterexamples.

In focusing on the use of thought experiments as counterexamples, we are interested in what Machery calls the *material use* of thought experiments. This is when we use thought experiments "not to discover the meaning of words or the semantic content of concepts of philosophical interest, but to understand their referents" (Machery, 2017, p. 16). The motivation for such use goes as follows. Theories of knowledge aim to explain what knowledge *is* (in a roughly metaphysical sense). So, if we take the Gettier cases to be a counterexample to such theories, the conclusions of Gettier cases should also concern what knowledge is not (and not what 'knowledge' might not mean). So, thought experiments concern hypothetical situations "most relevant to *the nature* of the phenomena under investigation" (Williamson, 2007, p. 206, emphasis added).[5]

The main question that the *use* of thought experiments gives rise to is: "how a merely hypothetical case can teach us anything interesting about the world and can even count as a counterexample to some (often well-established) theory" (Cohnitz & Häggqvist, 2018, p. 408). It seems rather strange that we should let the hypothetical situations that philosophers dream up refute previously accepted and supported philosophical theories – e.g., what knowledge is or is not.

This question has made some philosophers, as we will see in Chapter 10, "ill at ease with thought experiment. They [. . .] have doubts about how small-scale science fiction could prove anything" (Sorensen, 1992a, p. 15). Providing an epistemology of thought experiments that demystifies how we get justified beliefs on the basis of thought experiments would go a long way to ease philosophers. In particular, in line with our methodological remarks from Chapter 1, we are looking for a non-exceptionalist analysis of thought experiments.[6]

[4] It is controversial to what extent philosophers accepted the JTB analysis prior to Gettier (1963): see Shope (1983, pp. 12-19) and Dutant (2015).

[5] There is another interesting use of thought experiments, their *exploratory use*. This is when we use thought experiments to go beyond the boundaries of our 'standard' conception of things. Though interesting and of some epistemic use, the epistemic use is not one of *justification*. As a result of an exploratory thought experiment, one is not necessarily justified in believing the conclusion of the thought experiment, rather they are motivated to "investigate the boundary conditions of accepted theories" (Machery, 2017, p. 16). See Machery (2017, sec. 1.1.2) for a very clear discussion of this and other uses of thought experiments.

[6] Let me stress that in this chapter and the next, although we will be talking about cognitively

9.1.1 Two Non-Exceptionalist Theories of Thought Experiments

Two non-exceptionalist lines of research on thought experiments are currently prominent. Theorists on, what I will call, the **F-line** focus on the *form* and *content* of (the reasoning induced by) philosophical thought experiments. For example, they aim to provide a rational reconstruction of the 'Gettier-reasoning' that supports the standard 'Gettier judgement' (knowledge is not justified true belief) in response to Gettier cases. Williamson (2007) and Geddes (2017) give broadly non-exceptionalist developments of the F-line; appealing only to ordinary cognitive capacities whose nature and reliability is amenable to scientific (e.g. evolutionary) explanation.[7] Meanwhile, theorists on, what I will call, the **X-line** (e.g. Weinberg et al., 2001; Swain et al., 2008; Wright, 2010; Starmans & Friedman, 2012; Nagel et al., 2013; Turri, 2013; Machery, 2017) focus on the question of *robustness*: which philosophical thought experiments, if any, elicit judgements that are stable and uniform across population and presentation? Standard scientific tools (i.e. rigorous experimental design and analysis) are deployed to clarify and assess thought experiments' trustworthiness, again on the naturalistic assumption that thought experiments utilise *ordinary* judgement (lest folk surveys be rendered irrelevant).

Interestingly, F-liners and X-liners proceed from opposing inclinations. F-liners typically assume that prominent instances of thought experiments induce good reasoning, yielding knowledge in paradigm cases (see Williamson, 2007, ch. 6 on the Gettier case). Recovering this possibility is taken as a mark of an adequate reconstruction. X-liners, in contrast, assume a sceptical stance: it is a matter of (empirical) scrutiny whether thought experiments deserves their cherished status in the philosopher's tool kit. In Chapter 10, I will elaborately compare these two non-exceptionalistic lines. In this chapter, I will discuss a particular debate *within* the F-line; focusing on a specific aspect of Williamson's theory.

Williamson (2007) analyses the epistemic content of Gettier thought experiments along broadly non-exceptionalist lines as *counterfactual reasoning*. Williamson's analysis draws out critical aspects for an analysis of thought experiments. I will dub these the question of **Form**, of **Justification**, and of **Detail** (to be specified in the next section). Respondents to Williamson's analysis (from within the F-line) have focused on *the problem of deviant defeat*, rooted in *deviant realisations* of Gettier cases. This problem aims at **Form** and **Detail** of Williamson's analysis of thought experiments (Ichikawa & Jarvis, 2009; Malmgren, 2011; Geddes, 2017). All of Williamson's critics argue that his account of **Form** needs to be amended.[8]

plausible epistemologies of possibility in general, we will abstract away from the specific theories discussed in the previous two parts of this dissertation.

[7] Malmgren (2011) and Ichikawa & Jarvis (2009, 2012, 2013) are F-liners that accommodate rationalism.

[8] Ichikawa & Jarvis (2009, 2013) also argue that his account of **Detail** must be amended.

I will argue that the problems stemming from deviant realisations are worse than previously recognised. That is, the problem of deviant defeat applies very generally, also to, for example, the account of Geddes (2017), which is explicitly designed to overcome these issues, and even to analyses based on the *indicative conditional*. Moreover, I argue that there is another, often overlooked, *problem of deviant disagreement* that targets **Justification** and **Detail** of analyses of thought experiments. Therefore, I suggest an amendment of **Detail**, as otherwise general approaches to both **Form** and **Justification** have to be abandoned. Secondly, I argue that it is actually **Justification** that is the core of Williamson's account (as opposed to the tendency to focus on **Form**). I will argue that retaining this aspect of his account, and with the amended **Detail**, results in the possibility of *pluralism* with respect to **Form**: Williamson's own account of **Form** can be preserved, as well as other accounts.

I conclude with suggesting that there are advantages to adopting an *indicative conditional analysis* of thought experiments. In particular, it emphasises the fact that "thought experimentation typically involves several *modal* judgements" and that one of the main questions that needs to be addressed is "why are we entitled to these modal judgements?" (Cohnitz & Häggqvist, 2018, p. 406, original emphasis). In line with the work of this dissertation, this chapter emphasises the *philosophical* importance of the epistemology of possibility: judging whether hypothetical situations are possible or not is crucial for the use of philosophical thought experiments.

9.2 Form, Justification, Detail, and Deviance

In this section, we will get some preamble out of the way concerning the set up of Gettier thought experiments; some terminological issues; and the problem of deviant defeat. To that end, I start with the description of, what I call, a *bare Gettier case*:[9]

Bare Gettier Case: Paul Jones has excellent evidence that Smith owns a Ford: Smith has shown up every day at work with a Ford, Smith claims to own a Ford et cetera. Paul concludes on this basis that Smith owns a Ford. Paul then concludes that someone in the office owns a Ford. But Smith is driving a rental, and is too embarrassed to admit it. Coincidentally, Brown, another office-mate, owns a Ford. So Paul's belief is true.

From this bare Gettier case, one (usually) gets to the following conclusion through the following judgement:

Gettier judgement: Paul has a justified true belief that someone in the office owns a Ford, but this belief is not knowledge.

Gettier conclusion: Knowledge is not justified true belief.

[9] This particular case description is adapted from Malmgren (2011).

I take this as our standard set up of the Gettier thought experiment (henceforth 'Gettier') for this chapter. Moreover, I take this as a paradigmatic instance of a thought experiment that we are interested in from an epistemological point of view (see also Williamson, 2007; Machery, 2017; Hawke & Schoonen, 2020). As mentioned above, Williamson (2007) is interested in a rational reconstruction as the epistemology of thought experiments. I suggest that a satisfying rational reconstruction of Gettier should at least answer the following questions:[10]

- Question of **Form**: what is the correct (rational) reconstruction of how the agent is in a position to accept the Gettier conclusion?
- Question of **Justification**: what justifies the agent in accepting the premises (of the reconstruction)? Which cognitive capacities are the source of the justification?
- Question of **Detail**: how does the subject discern intended, critical features of a Gettier case that are not explicitly mentioned in the bare Gettier case (e.g., that Paul lacks knowledge)?

Below, I will elaborate in more detail Williamson's own answer to these questions. For now, let me simply stress that any account that aims to provide an analysis or rational reconstruction of thought experiments should at least say something about each of these questions.

The bare Gettier case described above is compatible and consistent with *deviant realisations*: realisations with additional details that are unintended by the experimenter and that undermine the Gettier conclusion. As Malmgren (2011) puts it,

> [A] standard case description is radically incomplete, and there may be ways of completing it on which the subject does *not* have a justified true belief without knowledge. [...] These realizations are *deviant*, [...]. That is, the corresponding interpretations of the case description are clearly *unintended*—the description was not meant to be read in some such way'. (pp. 275-277, original emphases)

In general, there are two ways in which a Gettier case can be a deviant realisation: x knows φ by some other means or x's justification is defeated. For example, in our bare Gettier case, Paul could know that Brown owns a Ford through additional testimony, thereby making the beliefs about Smith's owning a Ford irrelevant for the justified true belief to be knowledge. Conversely, Paul could be prone to hallucinating people driving Fords, this would undermine their justification for the belief that Smith owns a Ford, making it a situation where there is no justified true belief to begin with. In both cases, the Gettier conclusion is undermined.

[10] These are distilled from Williamson's own work.

In order to remain clear on these issues throughout this chapter, I will use the following terminology:

- '*GC*' refers to *bare Gettier cases*;
- 'GC^+' refers to any *intended enrichment*; and
- 'GC^*' refers to any *deviant realisation*.

Deviant realisations give rise to three more challenges that a satisfactory analysis of (Gettier) thought experiments should capture:

Defeat: Intuitively, certain information can defeat Gettier reasoning and intuitively certain information *cannot* defeat Gettier reasoning.

Accidental Deviancy: A subject that considers only an unintended, deviant case when presented with *GC* is making a *mistake*, resulting in reasoning that is *irrelevant to* and *independent of* a successful thought experiment, even if cogent (e.g., a correct judgement about GC^* (e.g., there is knowledge) is independent from correct Gettier reasoning (e.g., there is no knowledge)).

Real Disagreement: It is possible for a subject to consider the *intended* GC^+ but nevertheless contradict the Gettier conclusion (see results from experimental philosophy, e.g., Machery, 2017). Such a subject *genuinely disagrees* with the standard (correct) evaluation of *GC*.

The problem most discussed in the literature is the problem of *deviant defeat*: deviant realisations seem to defeat the Gettier reasoning on a particular analysis, however, the deviant realisation, intuitively, *shouldn't* defeat the Gettier reasoning. As Geddes (2017) puts it, "our semantic intuitions seem to tell us that such 'deviant' instances of the case have no bearing on the correctness of the judgement" (p. 39).

With all these desiderata for a satisfactory analysis of (Gettier) thought experiments in place and the terminological preamble out of the way, let us now turn to Williamson's analysis and some of the virtues thereof.

9.3 Williamson's Analysis of Gettier Reasoning

In this section, I will spell out Williamson's (2007) analysis of Gettier. I will go through his analysis of **Form**, **Justification**, and **Detail** in turn and then I will discuss some advantages of such a Williamsonian analysis. That is, I will give some motivations for only a *minimal* departure in light of the problems to be raised in the next section.

Williamson (2007, ch. 6) provides the following analysis of how the subject comes to the Gettier judgement – i.e., of **Form**:[11]

[11] This is a highly simplified version of Williamson's original formulation. For instance,

W1 $\Diamond\exists x GC(x,p)$

W2 $\exists x GC(x,p) \mathrel{\Box\!\!\rightarrow} \exists x(JTB(x,p) \wedge \neg K(x,p))$

C1 $\Diamond\exists x(JTB(x,p) \wedge \neg K(x,p))$

C2 $\neg\Box(JTB \leftrightarrow K)$

'$GC(x,p)$' means 'x stands in the relevant 'Gettier' relation to the proposition p'. Similar analyses apply to '$K(x,p)$' and '$JTB(x,p)$'. This analysis of **Form** is the most famous and most criticised part of Williamson's account (Ichikawa, 2009; Ichikawa & Jarvis, 2009; Malmgren, 2011; Geddes, 2017). I will say more about it in the next section, when I will go on to discuss the objections raised against it. What is important to note from the get-go is that the modalities involved in this analysis are *objective* or *metaphysical* modalities (as opposed to, e.g., epistemic or deontic), resulting in the material use of Gettier cases.

Concerning **Justification**, Williamson (2016a) specifies (more explicitly than Williamson, 2007) what makes the subject justified in accepting the premises and what cognitive capacities underlie that justification. The first premise is justified by whatever justifies modest ordinary possibility claims. The subject's psychological state is that of belief in a possibility claim (e.g., $B(\Diamond\exists x GC(x,p))$). More importantly, the justification of the second premise is, roughly, whatever justifies ordinary conditional belief (e.g., $B(JTB \wedge \neg K \mid GC)$).[12] This means that if the Gettier case is actual, then the justification is just ordinary belief update, whereas if the case is hypothetical it is 'offline' belief update. The latter might, for example, be an exercise of *reality-oriented imagination* understood as simulated rational belief update (see also Chapter 4). That is,

> one supposes the antecedent and develops the supposition, adding further judgments within the supposition by reasoning, offline predictive mechanisms, and other offline judgments. The imagining may but need not be perceptual imagining. [...] To a first approximation: one asserts

Williamson's preferred analysis of **W2** is:

$$\exists x\exists p[GC(x,p) \mathrel{\Box\!\!\rightarrow} (\forall x\forall p[GC(x,p) \supset (JTB(x,p) \wedge \neg K(x,p))])]$$

There are two reasons why you might prefer this formalisation: (i) this seems to best deal with complications of the 'Donkey-anaphora' kind (see Williamson, 2007, ch. 6.4) and (ii) this seems to best capture the fact that people that do not draw the conclusion that there is JTB without knowledge actually disagree. I use this simplification for ease of exposition, with it, I intend to refer to Williamson's original analysis.

[12] The 'roughly' is important, for there are some important subtleties between the justification for accepting conditional beliefs and the justification for accepting counterfactuals and I do not want to suggest that Williamson conflates the two (he is actually very explicit about the difference, see Williamson, 2016a).

> the counterfactual conditional if and only if the development eventually leads one to add the consequent. (Williamson, 2007, pp. 152-153)

This appeal to reality-oriented imagination, in epistemological support of counterfactual and mere possibility claims, is defended by cognitive scientists (see references in footnote 13).

Finally, for **Detail**, the 'counterfactual' in Williamson's counterfactual-analysis plays an important role. "By using the counterfactual conditional, we in effect leave the world to fill in the details of the story, rather than trying to do it all ourselves" (2007, p. 186). What this results in is that the relevant enrichment of GC, i.e. GC^+, is determined by what the nearest GC-worlds are like (Lewis, 1973b). Similarly, the subject's expectations of what this enrichment is like, is determined (again, roughly) by which GC-worlds they think are most plausible.

The above Williamsonian analysis has a number of advantages that are worth emphasising before we turn to discuss the recent objections against this analysis.

- *Fit with pre-theory.* First of all, the account aligns nicely with a pre-theoretic description of participating in a Gettier thought experiment: the given text is a springboard for imagining a scenario that one judges to have certain epistemic features, leading to a conclusion about the nature of knowledge.

- *Non-Exceptionalistic.* Reality-oriented imagination as simulation and as part of the epistemology of counterfactuals is taken seriously in cognitive science. Moreover, counterfactual and possibility claims are ordinary phenomena subject to evolutionary explanation.[13]

- *Allows for real disagreement.* **W2** is inconsistent with **W2*** where '$\neg K$' is replaced with 'K', or 'JTB' with '$\neg JTB$' (given that the antecedent of **W2/W2*** is possible). This captures the idea that if people draw the conclusion that there is justified true belief *and* knowledge, they would actually disagree with someone who draws the 'right' conclusion.

- *Plausibly cogent reasoning.* The argument is valid. Scepticism about **Form** or **Justification** amounts to scepticism about ordinary faculties for judging ordinary possibility and counterfactual claims and scepticism about reliability of conditional belief concerning ordinary subject matter. Yet, the reasoning is predicted to be fallible. That is, there is no infallible 'modal vision' or 'rationalistic intuition' or anything of that sort involved (e.g. Bealer, 2000). This

[13] Here is a small sample of references: Hesslow (2002); Byrne (2005); Goldman (2006); Nichols (2006a); Epstude & Roese (2008); Rafetseder et al. (2010); De Brigard et al. (2013); Gopnik & Walker (2013); Walker & Gopnik (2013); Byrne (2016); Lane et al. (2016) and the references therein.

leaves room for one to be sceptical about the essence of far-fetched cases (bearing on **W1**) or lack of conceptual competence/requisite background knowledge (bearing on **W1**, **W2**). Compare **Gettier** to, for example, the zombie thought experiment or an experiment where one puts on hold the laws of nature. As Williamson points out "we are more reliable in evaluating some kinds [of counterfactuals] than others. [. . .] We may be correspondingly more reliable in evaluating possibility of everyday scenarios than of 'far-our' ones, *and extra caution may be called for in the latter case*" (2007, p. 164, emphasis added). This leaves room for *modal modesty*, in the spirit of Van Inwagen (1998) and Hawke (2011) (see also Chapter 11). As Strohminger and Yli-Vakkuri put it when discussing moderate modal scepticism in a Williamsonian epistemology of modality:

> it seems plausible to us—although Williamson does not say this—that, at least generically speaking, the more distant a state of affairs p is from actuality, the more difficult it will be to imagine how things would be if p were to obtain in the amount of detail required for knowing that p does not counterfactually imply a contradiction.
> (2018b, p. 315)

- *Substantial subject matter.* Knowledge itself is at issue, not just the concept of knowledge: metaphysical not conceptual modalities in play. That is, the thought experiment is used in *material-mode* reasoning (Machery, 2017). This is relevant in explaining that the thought experiment results in knowledge of the nature of knowledge, as opposed to the concept of knowledge as we use it.[14]

- *Accommodates actual and hypothetical cases.* The account is broad enough to explain why there is nothing essentially fictional about Gettier reasoning (e.g., see Williamson's 2007 actualised Gettier cases).

9.4 Problem of Deviant Defeat

Having described Williamson's analysis of thought experiments and the benefits thereof, let us now turn to some of the criticism of his approach. As mentioned above, most of the critics focus on Williamson's analysis of **Form** and on *the problem of deviant defeat*. In order to properly spell out what the core features are of this problem, I will present a slightly abstracted form of Williamson's analysis.

[14] It is unclear whether Machery (2017) would accept this as a virtue, as he forcefully argues for the use of naturalised conceptual analysis. On the other hand, it seems that if one can defend a plausible naturalistic analysis of the method of cases, Machery's worries might subside (see Chapter 10 and Hawke & Schoonen, 2020).

Note that Williamson's analysis of **Form** is a particular instance of a *(restricted) conditional analysis* of Gettier reasoning. That is, we can represent the analysis in an abstract and non-committal (about the particular conditional) way:[15]

P1 $\Diamond GC$

P2 $GC \Rightarrow (JTB \wedge \neg K)$

C1 $\Diamond(JTB \wedge \neg K)$

Where '$\varphi \Rightarrow \psi$' is any *restricted conditional*, acceptable only if: the *relevant* φ-worlds are all ψ-worlds.[16] Examples of how relevance is determined are one or more of the following features: (i) objective features of modal space, (ii) the actual world, (iii) the subject's knowledge of actuality, or (iv) φ's subject matter.

We are now in a position to specify the problem of deviant defeat through the notion of *defiant information*. Defiant information is information that shows that some *relevant* GC-worlds are GC^*-worlds. That is, information showing that some relevant worlds are deviant realisations. For example, if it turns out that there is a relevant world where Paul is told that Brown owns a Ford, this would defeat the Gettier judgement and, accordingly, the Gettier conclusion. This is problematic when the defiant information intuitively is *irrelevant* to the thought experiment and thus, intuitively, *shouldn't* defeat the Gettier reasoning.

Below I will discuss the particular instances of the problem of deviant defeat. First, I will recite the problem for Williamson's account (which is raised by, amongst others, Ichikawa & Jarvis, 2009; Malmgren, 2011; and Geddes, 2017). After this, I will discuss similar problems for Geddes' own account – which is specifically designed to avoid it – and a roughly Ramseyan-indicative conditional approach. Showing that all such accounts are susceptible to the problem of deviant defeat is meant to show that the problem is much deeper than previously thought.

Problem of Deviant Defeat: Williamson

Remember that the crucial premise in Williamson's (2007) analysis of **Form** is his **W2**, which uses a counterfactual conditional. Given some standard assumptions about the meaning and semantics of counterfactuals, a counterfactual claim is defeated if there is at least one *closest* world that is a deviant case. That is, what

[15] From here on out, I'm going to represent the analysis of **Form** even more crudely. Instead of writing '$\exists x GC(x,p)$' for 'there is someone who stands in a relation as described by the Gettier case to a proposition', I will simply write 'GC' to denote the set of worlds in which this is the case. So, when I speak of a 'GC-world', I talk about a world where there is somebody who stands into a relation as described by the Gettier case to a proposition. Similarly for the other, e.g., JTB, K, etc., notational abbreviations. I will also ignore the final move to **C2**, which remains unchanged.

[16] This use of '$\Rightarrow$' is not to be confused with the way the symbol was used in Chapter 4, where I used it to refer to *the material conditional*.

determines relevance for the counterfactual conditional is closeness to the actual world, so defeating information is finding out that one of these closest worlds is a GC^*-world.

Now, it is rather easy to come up with an example where, in the actual world, I have a colleague that stands in the relevant, GC, way to a proposition that someone in their office owns a Ford. However, they are also aware that they are prone to hallucinate people driving Fords and prone to misremember people driving Fords. It seems then that, in the actual world, my colleague does *not* have a justified true belief without knowledge that someone in their office owns a Ford (example adapted from Malmgren, 2011). On the assumption that the actual world is always amongst the closest worlds to actuality, this case would defeat the formal argument captured by Williamson's analysis of **Form**. However, intuitively, my colleague and their proneness to hallucinations is completely irrelevant to the Gettier judgement and *shouldn't* defeat the Gettier reasoning. This, according to many, is problematic for Williamson's account (Ichikawa & Jarvis, 2009; Malmgren, 2011; Geddes, 2017).

Problem of Deviant Defeat: Geddes

Geddes (2017) sticks closest to Williamson's original analysis in light of the objection and is thus, arguably, able to reap the benefits of Williamson's account if unaffected by the objections (although Geddes is never explicit about these issues). What Geddes tries to do is capitalise on the fact that these deviant cases are *deviant* or *abnormal.* He argues that it is clear that when we are thinking about thought experiment cases, we are not concerned with these abnormal cases, but we are concerned with the *normal* ones. He proposes that "our judgements about thought experiments are typically judgements of just this sort—in other words, that they are typically judgements about what hypothetical scenarios *normally counterfactually suffice* for" (2017, p. 45, original emphasis). So, Geddes suggests to alter the counterfactual analysis as follows:

$\blacksquare\!\!\rightarrow$ "If someone were to stand to a proposition p as in the Gettier Case, then, normally, he would have a justified true belief that p but not know that p" (p. 48)

If we now turn to the problem of deviant defeat and analyse Geddes' account along the restricted conditional analysis, we note that for Geddes the *relevant* φ-worlds are the *closest normal worlds* where GC holds.[17] Again, it might be the case that there

[17] It is unclear whether Geddes (2017, pp. 49-50) intends to be taken in this way as he explicitly refrains from providing a precise semantic analysis. Another interpretation might be that in the relevant subset of worlds, *most* of those GC worlds are worlds where there is JTB without knowledge. I think that even if he means this, his account would still be susceptible to deviation. However, I will stick with the above interpretation of his conditional as this way of reformulating his gloss brings out the objectivity of his notion, the objectivity that is reflected in the features

is a closest normal world where there is JTB without knowledge (Geddes himself acknowledges that we might be wrong about what is normal).[18] In this case, Geddes' analysis would predict the Gettier reasoning to be defeated, but our intuitions suggest that the fact that all closest normal worlds are deviant is irrelevant and that these *shouldn't* defeat the Gettier reasoning. This, analogous to the objection to Williamson's account, is a problem for Geddes' amendment.

Problem of Deviant Defeat: Ramseyan Indicative

Both Williamson's (2007) and Geddes' (2017) account of **Form** are susceptible to the problem of deviant defeat due to deviant realisations. One might think that this is the case because both analyses rely on *objective* features of modal space. That is, both accounts rely on features of modal space that we can be mistaken about: Williamson focuses on objective features of actuality, whereas Geddes focuses more broadly on objective features of modal space. So, it is natural to wonder whether an account that is more *subjective* would avoid these troubles.

One intuitive option would be to replace the counterfactual with the indicative conditional. Consider, for example, the following description of how we come to know whether an indicative conditional is true.

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.
>
> (Stalnaker, 1968, p. 102)

Geddes ascribes to normality: (i) is not a matter of what happens in the actual instance, (ii) we can be wrong about normality, and (iii) our interests do not drive what is normal.

[18] One might worry that it seems strange that there can be *normal* worlds that are worlds where there is deviation. It is crucial here to understand what I mean; what is meant is, as Malmgren (2011) notes, that the realised scenario does not match our *intention* of getting to a situation where there is JTB without knowledge. Now, to understand how a normal world can be deviant is as follows: according to Geddes the actual world need not be normal, so, for all we know, we live in an abnormal world (in this case with regards to what we take JTB to be). Maybe, it is the case that, influenced by the abnormal world we live in, we think that knowledge is not justified true belief. However, maybe, it is in fact normal that knowledge *is* justified true belief. So, in the normal worlds there is either not-JTB or there is knowledge. But, in *our* sense, these are deviant realisations.

Geddes, in a sense, replies to this worry when he notes that "it surely does not seem plausible to us that we inhabit a world in which [normality is] in fact vastly different from what we take [it] to be" (2017, p. 50). However, one can sketch a model on which we cannot rule out that normality is vastly different from what we take it to be. The general point here is that in order to determine what is normal on Geddes' account would be to know what modal space looks like and this is something that seems highly unlikely that we can do. To merely assume that certain things are normal (or that modal space is such that certain things come out as normal), would be to beg the question.

This, roughly Ramseyan, analysis of the indicative conditional sounds remarkably like Williamson's (2007, ch. 5) story concerning the epistemology of counterfactuals (Stalnaker, 1968, holds a 'closest possible worlds' analysis, related to belief updates, for both the indicative and counterfactual conditionals).[19] This then is, *prima facie*, a suitable solution for an account of **Form**. One would instantiate **P2** as follows:

I2 $GC \rightarrow (JTB \wedge \neg K)$

When one thinks about the indicative along these, Ramseyan, lines, you also take it to be a restricted conditional of sorts. Namely, we only look at those GC worlds that are most plausible relative to what the agent knows about actuality. Note that in this case, the analysis is much more 'subjective' than the objective (normal) counterfactual analyses.

However, think about when an indicative conditional is, usually, taken to be false: when the antecedent is true and the consequent is false.[20] Reasoning according to this indicative analysis is defeated if the actual world is a (known) GC world where there is either no justified true belief or there is knowledge. Note that this means that the reasoning on such an indicative analysis is thus defeated by *actual deviant realisation* (similar to Williamson's original account). Yet, per the usual scheme, such an actual deviant realisation should not, intuitively, defeat the Gettier reasoning. So, given that the Ramseyan indicative analysis predicts that it does, such an analysis also falls victim to the problem of deviant defeat.

9.5 Problems of Disagreement

Above I showed that the problem of deviant defeat is much deeper and widespread than previously thought. It does not only affect Williamson's original account, but also Geddes' account, which was specifically designed to avoid it, and a more subjective Ramseyan indicative conditional account. Before I discuss my proposed solution in the next section, let me briefly discuss one final option. That is, one might think that given the troubles spelled by deviant realisations, we should maybe give up accounts that involve instances of a restricted condition. One such radically different account of **Form** is that of Malmgren (2011).

Malmgren suggests that the correct rational reconstruction of Gettier reasoning should take the following form:

[19] In more recent work, Williamson (2016a) explicitly mentions that the epistemology of indicative conditionals and counterfactuals are very similar, modulo some subtle differences in the cognitive processes.

[20] Even those who think that indicative conditionals *do not* express propositions generally agree with this (e.g., Edgington, 1995; Bennett, 2003).

M1 $\Diamond(GC \wedge JTB \wedge \neg K)$

C1 $\neg\Box(JTB \leftrightarrow K)$

There are many interesting things to say about Malmgren's account (for example, about her rationalistic justification for the acceptance of **M1**). However, here I want to focus on the problems of *accidental deviancy* and *real disagreement*. In this section I will first elaborate a bit more on these two problems and then address why I take the latter to block the retreat to Malmgren's account.

It seems a rather trivial point that sometimes people who are told a version of the Gettier thought experiment do not draw the Gettier conclusion. Let's call *Gettier misjudgement* the judgements of x's having neither justified true belief nor knowledge (or x's having justified true belief *and* knowledge). Given that we have pulled apart **Detail** and **Justification**, we can provide a more detailed analysis of what goes on in the case of such Gettier misjudgements. It seems that there are two distinct ways in which a Gettier misjudgement might arise.

- The agent fills in some *unintended* epistemically relevant features.
- The agent *does* fill out the scenario as intended, but they draw the conclusion that there is no justified true belief or that there is knowledge.

Satisfactory reconstructions of what goes on when we reason about thought experiments should be able to account for the fact that the misjudgement in the former case is *irrelevant*, whereas in the latter case the subject *genuinely* disagrees with the Gettier judgement. The former is the problem of *accidental deviancy* and the latter the problem of *real disagreement*.

I will first discuss the problem of accidental deviancy as an additional, novel problem for Williamson's account, before I turn to discuss Malmgren's account of **Form**, which is susceptible to the problem of real disagreement.

9.5.1 Problem of Accidental Deviancy

Arguably, a lot of the misjudgements are of the 'accidental deviancy' kind. These are cases where the subject has, accidentally, filled in some details in a deviant way. An initial response to such misjudgements would be to ask why the agent thinks that there is no justified true belief or knowledge. The odds are that they have failed to fill in an epistemically relevant intended feature or that they have filled in something that was not intended. What one should do in such a situation is correct the agent and provide them, explicitly, with the information that they failed to (or incorrectly) apprehend(ed). Williamson (2007) points to something similar when he says that "even when lacunae are identified in a thought experiment, the most likely response in practice is just to *add further stipulations to the specification of the case*, [...], so as to preserve the original structure of the argument" (p. 204, emphasis added).

Even though Williamson seems to acknowledge that accidental deviancy should be seen as a *mistake* on the subject's side and should be *irrelevant* and *independent* of successful Gettier reasoning, his analysis seems to be unable to accommodate this intuition. Consider the story that Williamson tells concerning **Justification**: subjects that evaluate the intended realisation of the Gettier thought experiment are in the psychological state of accepting JTB without knowledge conditional on GC. More formally,

$$B(JTB \wedge \neg K \mid GC)$$

According to this story, a subject that, against the experimenter's intentions, makes a mistake and ends up believing an accidental deviant realisation is in the following belief-state (if the deviant realisation concerns there being knowledge as opposed to there not being JTB):

$$B(JTB \wedge K \mid GC)$$

Note that these are incompatible mental states. So *accidental* evaluation of GC^* is neither irrelevant nor independent of reasoning in the successful experiment on this analysis. That is, by Williamsonian lights, the subject *is* reasoning as in a successful experiment and the result is *incompatible* with reasoning with GC^+. This is, I suggest, counter-intuitive: when a subject accidentally misinterprets the thought experiment, one should not conclude that they reasoned successfully. Mistakes of this kind seem to be completely irrelevant to successful Gettier-reasoning. This poses an additional, novel problem for Williamson's account.

Having made this side-comment, let us now return to the problem of real disagreement that, I suggest, blocks a retreat to Malmgren's analysis of **Form**.

9.5.2 Problem of Real Disagreement

The problem of real disagreement concerns cases where the subject does get the intended enrichment of the Gettier case, but still holds there is either no justified true belief or that there is in fact knowledge.[21] We would like our analysis to reflect the intuition that in such cases there is real disagreement. Recent work in experimental philosophy shows that this is not merely a theoretical issue, but that people do actually have Gettier misjudgements (under certain presentations of

[21] One thing to note that is interesting concerns the relation between this problem of disagreement and the account of Grundmann & Horvath (2014). They argue that there is an interpretation of the Gettier-case that is 'deviant realisations proof' and, more importantly, that this is the interpretation that most of us (and especially epistemologists) get to. It seems then, that in order to accommodate the data of Machery (2017) (i.e., that there is some, albeit small, disagreement in the judgement) they have to argue that this is *not* due to a misunderstanding, but that it is always a case of genuine disagreement. This, to me, seems rather strong.

the Gettier case) (see Machery, 2017; Machery et al., 2018a). Malmgren's (2011) account of thought experiment reasoning (repeated below), however, seems unable to capture this phenomenon.

M1 $\lozenge(GC \wedge JTB \wedge \neg K)$

C1 $\neg\Box(JTB \leftrightarrow K)$

Accordingly, the analysis of a *mis*judgement on her account would be the following

M1* $\lozenge(GC \wedge JTB \wedge K)$ (alternatively, '$\lozenge(GC \wedge \neg JTB \wedge \neg K)$')

However, **M1** and **M1*** are jointly consistent; they can even be both true at the same time. This means that, on Malmgren's analysis, there is no way to capture the fact that misjudgements, based on intended enrichments, are symptomatic of real disagreement.[22]

I suggest that when there is a Gettier misjudgement, one of two things could have occurred: (1) something has gone wrong when the subject filled in the epistemically relevant details or (2) the subject has gotten the intended enrichment, yet genuinely disagrees with the judgement. Looking at these two kinds of error, I think that one does best to avoid the error at the level of **Detail**. However, disagreement at the level of **Justification** can be seen as *pure* disagreement about which judgement there is to be drawn. Given that Malmgren's (2011) account is unable to account for the genuine disagreement, I take it that her suggested account of **Form** is not a viable option as an analysis of thought experiments that avoids the problem of deviant defeat.

9.6 Intended Enrichment as Input

In Section 9.4, I argued that the problem of deviant defeat affects a much larger group of theories than just Williamson's (2007) theory. I take this to show that the problem is worse than recognised and proposed solutions, such as that of Geddes (2017), do not avoid the issues. Secondly, in Section 9.5, I showed that a retreat to a radically different account of **Form**, such as that of Malmgren (2011), raises a different problem, namely that of real disagreement. In this section, I will elaborate on my positive proposal. I aim to retain Williamson's account of **Justification** and do so by suggesting a different analysis of **Detail**. With this amended account of **Detail** and Williamson's account of **Justification** retained, the precise nature of the

[22] Geddes (2017) also accuses Malmgren of being unable to analyse *misjudgements*. However, he only notes this particular instance of Malmgren's account, whereas I spell it out more broadly as a general requirement.

conditional in the second premise of **Form** becomes immaterial. This opens up the possibility of *pluralism* with respect to **Form**.

I will first set out my positive proposal and then address an intuitive objection that one might have. Secondly, I will compare my suggestion to and distance myself from the proposal of Ichikawa & Jarvis (2009), who also suggest to amend the analysis of **Detail**. Then, in the next section, I will turn to why this allows the possibility of pluralism with respect to **Form** and make a tentative suggestion why a Ramseyan indicative account of **Form** might have a slight preference over the Williamsonian counterfactual analysis.

As we saw above, the problem of deviancy is widespread and hard to escape. However, the problems disappear if the reconstruction is revised along the following lines. Instead of taking the bare Gettier case, GC, as input, we reconstruct **Form** as having the *intended enrichment*, GC^+, as input.[23] The resulting analysis of **Form** is given below:

R1 $\Diamond GC^+$

R2 $GC^+ \Rightarrow (JTB \wedge \neg K)$

C1 $\Diamond(JTB \wedge \neg K)$

Note that I have not used any particular instantiation of the restricted strict conditional. This is because taking the intended enrichment as input allows one to use a range of different conditionals; I will discuss this in more detail in the next section. For now, it suffices to note that this analysis has the advantage that it is not susceptible to the problems of deviancy. This is because, by definition, taking GC^+ as the input rules out all the deviant realisations.

I intend to retain Williamson's (2007; 2016a) account of **Justification**. Most of the literature engaging with Williamson's (2007) general analysis of thought experiments has focused on his account of **Form**, thereby ignoring what I take to be a major contribution in its own right, his account of **Justification**. Taking the intended enrichment as input does not affect the story that Williamson tells with regards to **Justification** and we can simply retain it as is. So, on this analysis the underlying mental states are, for **R1**, $B(\Diamond GC^+)$ and, for **R2**, $B(JTB \wedge \neg K \mid GC^+)$. Note that this has the nice effect that, on this account, substituting GC^* for GC^+ issues *irrelevant* and *independent* reasoning. That is, we avoid the issues concerning accidental deviancy.

[23] Note that the terminology being used here is a bit too crude. In the original Williamsonian analysis, the account of **Form** is: '$\Diamond \exists x \exists p[GC(x, p)]$'. This reads, 'It is possible that a person stands in a relation to a proposition as described by the bare Gettier case'. So, the proposed revision should be read as, 'it is possible that a person stands in a relation to a proposition as described by the intended enrichment of the bare Gettier case'. I take the crudeness in the main text to be shorthand for this more complicated expression.

With this particular account of **Form** and our retaining Williamson's analysis of **Justification**, we depart, in a minimal way, from Williamson's original account with respect to the analysis of **Detail**. Remember that Williamson's use of the counterfactual allowed him to leave it to the world "to fill in the details of the story, rather than trying to do it all ourselves" (2007, p. 186). However, on my account the enrichment is no longer folded into (i) acceptability conditions of $\Rightarrow$ and (ii) the subject's expectations about GC-worlds. Rather, the enrichment is determined by the *experimenter's intentions* and communicated to the subject using ordinary conversational mechanisms.

This last thought might, *prima facie*, give rise to the following objection, an objection that, in some way, is already foreshadowed by Williamson himself. The problem, one might think, is that no experimenter or subject can explicitly enumerate or rule out *every possible* deviation, so how can they treat as given or communicate the intended enrichment that rules out all these possible deviations?

I agree that this is in fact an important, and even worrisome, issue that needs to be addressed. However, I take it that this is not a particular problem for my account of Gettier reasoning. Rather it is a general problem for theoreticians about communication, *independent* of capturing Gettier reasoning. Consider for example the following scenario:

> Someone is late for a meeting that was announced by email the previous day. Walking in, they say "Sorry, I just discovered the email was sent to my spam folder". You judge on this information, as the speaker intends, that, this morning, they were ignorant of the meeting's existence. But they did not explicitly rule out all of the other ways they could have come to know that there was a meeting (e.g. speaking to a colleague).

The question is, what was the implicit intended content and how was it communicated? In this scenario, the same problem arises as the potential worry for our account of Gettier reasoning. I take this to show that this problem does not affect my analysis of Gettier reasoning in particular.

9.6.1 Ichikawa & Jarvis: Enrichments as Fictions

Before we turn to our final section about pluralism with respect to **Form**, let me briefly discuss another account that proposes to amend the analysis of **Detail**: Ichikawa & Jarvis (2009). Ichikawa and Jarvis agree that the bare case is enriched independently of the Gettier reasoning (i.e., not folded into the acceptability conditions of $\Rightarrow$). Their claim is that the bare case is enriched by the same processes that enrich a bare fictional text into the 'full story'.[24] In particular, they suggest that we

[24] Their account is rather subtle and very detailed. I will present here a very crude version of their analysis as my reasons for distancing myself from them does not hinge on the details.

grasp a set of propositions, $\mathcal{P}$, that is true in exactly the same set of worlds as what is true 'according to the Gettier fiction'. Let me mention three, possibly worrisome, points that might persuade one to prefer my proposal over that of Ichikawa and Jarvis.

First of all, let me echo a worry made by Malmgren (2011), namely, that of impossible fictions (see Malmgren, 2011, pp. 304-306). One may worry about why the set of propositions denoted by an enriched fiction would itself be consistent (i.e., that all the propositions are compossible). Ichikawa & Jarvis (2009) do have some story to tell, where they suggest that through the use of conceptual rules, we can, eventually, get to know whether the set of propositions is (jointly) metaphysically possible. Their analysis is rather wanting, but I will focus here on one particular issue. The analysis of Ichikawa and Jarvis hinges, for the most part, on the *internal* consistency of the story, but it seems perfectly plausible that we can tell a consistent, impossible story (e.g., one about unicorns, or consistent, backward, time-travel tales, see also Priest, 1997). These are, seemingly, not ruled out by their analysis (see Ichikawa & Jarvis, 2009, pp. 234-235, for their account of internal consistency). Moreover, one might worry, as Malmgren does, that even if GC^+ is internally consistent, what guarantees us that all the imported background information is *compossible* with it?

Secondly, there is the issue of how one gets from GC to GC^+. Ichikawa and Jarvis tell us no story on how the details are filled in to get to the enrichment, other than that this is the same as how people fill in the details in fiction. Even though the similarity between fiction and thought experiments might be right in this aspect, this is still not an explanation of *how* this is done. There might be a way for Ichikawa and Jarvis to get around these first two worries by, like me, appealing to the author's intentions. However, their detour through the 'truth-in-fiction' analysis is a gratuitous complication that, I believe, is not needed. If anything, my analysis will have the benefits of their analysis without this unnecessary complication. A worry that remains in this vicinity, however, concerns their focus on the fictionalness of the Gettier cases. As Williamson (2007) shows, there are also actual instances of Gettier cases and it is not altogether clear if their account is able to deal with these.

A final worry concerns their account of **Justification**. Ichikawa and Jarvis propose the following account of **Form**:

IJ1 $\Diamond GC^+$

IJ2 $\Box[GC^+ \supset (JTB \wedge \neg K)]$

C1 $\Diamond(JTB \wedge \neg K)$

So, their proposed account of **Form** introduces a *necessity premise*. They suggest that the subject is justified in accepting this premise based on something like 'rationalistic insight' (Ichikawa & Jarvis, 2009, 2013). Accepting such a rationalistic account of **Justification** and **Form** might be problematic if one wants to pursue a non-exceptionalist analysis of thought experiments. I concur with the forceful arguments of, e.g., Williamson (2007) and Machery (2017) that we should in fact try to stay away from postulating such rationalistic faculties (see Chapter 1, Section 1.3).

Even though I agree with Ichikawa & Jarvis (2009) that the bare case description needs to be enriched independently of the Gettier reasoning, I think that there are a number of severe issues for their account. This is, *prima facie*, an advantage for my analysis and that of Williamson (2007), especially for those who have worries about rationalistic intuitions.

9.7 Pluralism and the Ramseyan Indicative

I've said very little about my account of **Form**. In particular, I have not said anything about how '$\Rightarrow$' in **R2** is supposed to be interpreted. In this section, I will elaborate on the hints towards the idea that with taking GC^+ as input, and the Williamsonian account of **Justification**, *pluralism* about **Form** seems possible.

In order to see why this is so, consider again what we said about variably strict conditionals: they are acceptable only if the relevant antecedent-worlds are all consequent-worlds. What we saw with the problem of deviant defeat was that most of the ways to define what 'relevant antecedent-worlds' are, did not rule out the possibility of deviant realisations being among the relevant antecedent-worlds. What my proposal comes down to is that we *only* look at relevant worlds where the intended enrichment is true (however these are determined). This way, we guarantee that all relevant worlds are intended enrichments. It should now be clear that whatever relevant subset of the intended enrichment-worlds one takes, the analysis works. We can either look at all of them, resulting in an analysis with a strict conditional; we can look at all the closest worlds, resulting in an analysis with the counterfactual conditional; we can look at all the closest normal worlds, as per Geddes' conditional; or we can look at all the most plausible worlds given the subjects knowledge, resulting in an analysis with the Ramseyan indicative conditional.

Accepting this amended version of **Detail** has consequences for the interpretation of the problem of deviant defeat. I have argued that it is in fact a deep and widespread problem, but the solution suggested here allows one to save Williamson's account from the problem (as well as others).

In the remainder of this section, I will focus on a different instantiation of the revised version, namely the Ramseyan indicative conditional version (as represented below). I will, very tentatively, suggest some light, but interesting, advantages of such an analysis over the Williamsonian analysis.

Ra $\Diamond GC^+$

Rb $GC^+ \rightarrow (JTB \wedge \neg K)$

C1 $\Diamond(JTB \wedge \neg K)$

Much of what I will say below concerns the story told about **Justification** in parallel with a Ramseyan analysis of **Form**. I will, therefore, spell out that story in a bit more detail. I take it that belief-states concern what the subject takes to be most plausible (given what they know about the actual world). So, adopting the Williamsonian account of **Justification**, the justification for **Ra** is whatever justifies beliefs about ordinary possibilities. The subject who accepts the first premise, is in the following belief-state:

$$B(\Diamond GC^{+})$$

The worlds that they hold to be the most plausible, given their information (i.e., the most plausible candidates for being the actual world), are worlds at which it is possible that GC^{+}. Secondly, the justification for **Rb** is whatever justifies ordinary conditional belief. The subject who accepts this premise is in the following belief state:

$$B(JTB \wedge \neg K \mid GC^{+})$$

That is, the most plausible GC^{+}-worlds, relative to the subject's information, are all $(JTB \wedge \neg K)$-worlds. This might, for example, be based on a simulated belief-update with the information 'GC^{+}'.[25]

Note that on the Ramseyan account it becomes clear that an additional story needs to be told about why the subject is justified in concluding $\Diamond(JTB \wedge \neg K)$, from accepting the first two premises. For Williamson (2007) it is less pressing to tell such a story as for him the argument is formally valid and we might assume that we are justified to accept the conclusion of formally valid arguments of which we are justified in accepting the premises. However, on the Ramseyan indicative analysis, the argument is not straightforwardly valid. I suggest that, even though the inference is not formally valid, the following story concerning the justification of the inference still holds. If it is rational for a subject to accept $\Diamond GC^{+}$ and it is rational for them to accept $(JTB \wedge \neg K)$ given GC^{+}, then it is also rational for that subject to accept $\Diamond(JTB \wedge \neg K)$. This seems particularly plausible when we take this kind of simulated belief-update as the kind of *reality-oriented imagination* that Williamson (2007, 2016a) has in mind. For, as he points out, this kind of imagination is only useful when it depends on what we know of the world. Therefore, it has a "tendency to use something like rules of deductive logic", making it that this kind of simulated belief-update is "quite generally truth-preserving" (Williamson, 2016a, p. 122; see also Chapter 4).

[25] See Chapter 4 for a detailed discussion of imagination as simulated belief revision.

9.7.1 Ramseyan Indicative: Some Advantages

One reason why one might have a slight preference for the Ramseyan analysis over the Williamsonian analysis is that the former has some interesting theoretical advantages stemming from a closer relationship between **Form** and **Justification**. That is, on the Ramseyan analysis, the statements in the argument of **Form** directly express the relevant cognitive states of **Justification**. Consider the following instances of **Form**: the indicative, '$\varphi \rightarrow \psi$', and the counterfactual, '$\varphi \mathrel{\Box\!\!\rightarrow} \psi$'. The former expresses that ψ is reasonable to believe given information that φ (i.e., we assume a roughly Ramseyan analysis). The counterfactual, on the other hand, does not straightforwardly express this. This becomes clear when we consider the following counterfactual expression:

(8) If Oswald had not shot Kennedy then Kennedy would not have been shot.

Acceptance of (8) does not express that one would come to believe that Kennedy wasn't shot given the information that Oswald didn't shoot him (see also Bennett, 2003, p. 30).

The benefit of such a close link between **Form** and **Justification** is that this focuses our attention on the epistemically most interesting questions. (One might think that it is not the Ramseyan analysis in particular that does so, but rather the possibility of pluralism with respect to **Form** that focuses our attention on interesting questions about **Justification**.) For example, one might wonder what it takes for **Ra** (i.e., $\Diamond GC^{+}$) to be acceptable for a subject. In order to understand how subjects get to accept the Gettier conclusion, we should thus focus on a proper epistemology for accepting the first premise – i.e., an epistemology of possibility. In particular, given that we are aiming for a non-exceptionalist account of thought experiments, we need a cognitively plausible epistemology of possibility (see Chapter 1, Section 1.3). This deserves and requires further research, irrespective of one's preferential analysis of **Form**. As the Ramseyan analysis tracks **Justification** more closely with its instantiation of **Form**, these questions rise more naturally to the surface. As Kung phrases it, "[i]f you are attracted to this [i.e., counterexample] role for thought experiments, then you need some explanation of how thought experiments tell us about genuine metaphysical possibilities" (2017, p. 135).

One of the main questions in the epistemology of thought experiments (or, to talk boldly with Williamson, in the epistemology of philosophy) is: how do we justifiably believe (or know) hypothetical situations to be possible. The findings of the previous two parts of this dissertation (in particular Chapters 5 and 8) suggest that, while relying on our ordinary cognitive capacities, we can come to justifiably believe some possibility claims. If we can use these methods to also justifiably believe philosophically interesting possibilities, then the case for non-exceptionalist analyses of thought experiments that capture the fact that reasoning by thought experiments yields knowledge would be strengthened.

Experimental philosophers of the X-line, mentioned in the introduction to this chapter, are sceptical about the prospects of such a non-exceptionalist epistemology of thought experiments. They argue that philosophical cases have certain properties that make it so that our ordinary cognitive capacities do *not* produce reliable judgements about philosophically interesting possibilities. They suggest that we should therefore *shelve* the use of thought experiments completely. I will discuss some of the main arguments they present in favour of *scepticism* with regards to the use of thought experiments in the next chapter.

Chapter 10
Are Gettier Cases Disturbing?

> *[A]nalytic philosophy has entered a phase of systematic reassessment of what it was previously, and often uncritically, taken to be its standard methodology*
>
> – Angelucci & Arcangeli, 2019

In Chapter 9 we concluded that the reliance on thought experiments largely hinges on the question of whether we can justifiably believe a situation to be possible or not. Over the course of the dissertation (i.e., Chapters 3 - 8), we saw that there are ways through which we can acquire justified beliefs about what is possible and ways to extend this knowledge. If we could use these methods to gain knowledge of philosophically interesting possibilities, the Williamsonian F-line analysis of the method of cases we discussed in Chapter 9 would be strengthened. However, philosophers of the X-line are more sceptical. In particular, Machery (2017) argues for *radical restrictionism* and suggests that we should *abandon* the use of thought experiments in philosophy. In this chapter,[1] I will argue that Machery's severe conclusions are unfounded and that there are at least some instances of philosophically interesting thought experiments where we can justifiably come to believe the relevant possibility claims.

[1] The material of this Chapter is based on Hawke & Schoonen (2020).

10.1 Two Takes on Thought Experiments

Remember from Chapter 9 that we noted that there are two currently prominent non-exceptionalist lines of research on the method of cases (MoC).[2] Theorists on the **F-line** focus on the *form* and *content* of (the reasoning induced by) philosophical thought experiments. Theorists on the **X-line** (e.g. Weinberg et al., 2001; Swain et al., 2008; Wright, 2010; Starmans & Friedman, 2012; Nagel et al., 2013; Turri, 2013; Machery, 2017) focus on the question of *robustness*: which philosophical thought experiments, if any, elicit judgements that are stable and uniform across population and presentation?

These theorists proceed from opposing inclinations. F-liners typically assume that prominent instances of MoC induce good reasoning, yielding knowledge in paradigm cases. X-liners, on the other hand, assume a sceptical stance: it is a matter of (empirical) scrutiny whether MoC deserves its cherished status in the philosopher's tool kit. Machery (2017, pp. 6-8), for example, advocates *radical restrictionism*: in light of its empirically confirmed unreliability, traditional MoC should effectively be shelved and judgement about standard philosophical cases suspended, Gettier cases included.[3] In contrast, some prominent X-liners only endorse *moderate restrictionism* (Weinberg, 2007; Alexander & Weinberg, 2014): existing empirical results don't establish the widespread unreliability of philosophical thought experiments, but show that identifying trustworthy instances is a non-trivial empirical task.

The broad aim of this chapter is to clarify the interaction between the F-line and X-line, and gesture at the common path forward for non-exceptionalists. The narrow aim is to explore, in particular, how F-liner Williamson (2007) and X-liner Machery (2017) complement and contrast with each other. I will identify crucial shared commitments; rule on a disagreement about the force of the Gettier thought experiment (henceforth: Gettier); and thereby examine how far Williamsonians should accept radical Macherian conclusions.

[2] Following Machery (2017), the main antagonist of this chapter, I will use the phrase 'the method of cases' more often than before. As Machery (2017, p. 4) puts it, the method of cases consists of considering "actual or hypothetical situations (described by cases) and determine what facts hold in these situations," which is similar to how we described thought experiments in the previous chapter.

[3] Machery (2017, ch. 4) develops a second argument for abandoning MoC, as follows. Experimental investigation reveals that philosophical thought experiments yield inconsistent judgements among epistemic peers. If the disagreement is real, philosophers ought to suspend belief on the deliverances of MoC, lest they be dogmatic. But perhaps the disagreement is *merely* apparent: philosophers and non-philosophers differ in their interpretation of the cases. In this case, philosophers ought to focus on which interpretation reflects the most significant issues, to avoid over-emphasising merely parochial concerns. Concerning Gettier thought experiments, one can respond as follows: as I shall discuss, experimental results do *not* indicate robust disagreement among epistemic peers on the status of (certain) Gettier cases (Machery, 2017, sec. 4.1.4 explicitly concedes this). So Machery's dilemma plausibly doesn't get off the ground.

We proceed as follows. Section 10.2 isolates common ground between the Williamsonian F-line and the Macherian X-line. Section 10.3 uses it to explicate and criticise a Macherian case for pessimism about Gettier. Section 10.3.1 argues that Macherian pessimism hinges on the claim that Gettier cases have intrinsic features that disturb ordinarily reliable judgement. Section 10.4 argues that key Gettier cases are *not* disturbing. Section 10.5 considers implications for central arguments in Machery's *Philosophy Within Its Proper Bounds* (henceforth *PwPB*) and argues that Machery's argument for radical restrictionism is undermined if Gettier can paradigmatically be taken as reliable. Finally, I will present a cautious variant of Machery's argument, in support of a potent modal modesty that limits philosophy's theoretical ambitions, despite some preservation of traditional MoC.

10.2 Common Ground

Remember that on a broadly Williamsonian approach, a successful account of MoC has three features. First, it is non-exceptionalist. Second, it paints MoC as delivering (what Machery calls) *material-mode* conclusions: MoC, it is held, is not used "to discover the meaning of words or the semantic content of concepts of philosophical interest, but to understand their referents" (Machery, 2017, p. 16). Relatedly, applications of MoC are taken to establish metaphysical possibilities, the "sort of possibility most relevant to the nature of the phenomena under investigation" (Williamson, 2007, p. 206). Third, the account must explain the paradigmatic success of Gettier, on the hypothesis that "if any thought experiment can succeed in philosophy, then [Gettier's] do" (Williamson, 2007, p. 178).

Points of affinity with Machery (2017) are immediate. Machery agrees that MoC is best described as non-exceptionalist: the induced judgements "are warranted, if they are, for the very reason that everyday judgments are warranted, whatever that is" (Machery, 2017, p. 21). He agrees that MoC is best characterised as in the material-mode (2017, p. 16). Though cautious in his conclusions, he agrees that Gettier stands out as particularly robust: the judgements elicited by Gettier cases have only negligible demographic variation (Machery et al., 2018a) and only small to moderate ordering and framing effects (see Table 2.9 on pp. 86-87 of Machery, 2017). The folk apparently judge in accord with philosophical orthodoxy at a similar rate to their judgement of ignorance in response to a trivial 'false belief' case. This contrasts with early experimental studies that concluded significant demographic variation in judgement (Weinberg et al., 2001), but used small sample sizes and failed to be replicated (Nagel et al., 2013; Turri, 2013; Kim & Yuan, 2015; Sayadsayamdost, 2015). Indeed, Machery et al. (2018a) hypothesise that the Gettier judgement reflects universal features of folk epistemology (Machery et al., 2017 are more cautious).

To elaborate, consider a key Gettier case:

> **Hospital.** Paul Jones was worried because it was 10 pm and his wife Mary was not home from work yet. Usually she is home by 6 pm. He tried her cell phone but just kept getting her voicemail. Starting to worry that something might have happened to her, he decided to call some local hospitals to ask whether any patient by the name of "Mary Jones" had been admitted that evening. At the University Hospital, the person who answered his call confirmed that someone by that name had been admitted with major but not life-threatening injuries following a car crash. Paul grabbed his coat and rushed out to drive to University Hospital. As it turned out, the patient at University Hospital was not Paul's wife, but another woman with the same name. In fact, Paul's wife had a heart attack as she was leaving work, and was actually receiving treatment in Metropolitan Hospital, a few miles away.

Philosophical orthodoxy takes Hospital to induce the judgement that Paul has a justified true belief (his wife is in hospital) that isn't knowledge. Call this a *singular Gettier judgement*, supporting the *Gettier conclusion*: knowledge is not justified true belief. As it is tricky to explain precisely why Paul lacks knowledge (Shope, 1983), suggestive but non-committal terminology will be useful: Paul's belief is not knowledge since its grounds are not suitably *sensitive* to what makes it true – its truth is somehow *lucky*.

Credible studies indicate that Hospital induces widespread convergence on the singular Gettier judgement, bolstering philosophical orthodoxy.[4] Surveying over 2000 participants, Machery et al. (2017) found both men and women made the singular Gettier judgement at a rate of about 80%. Participants across 23 countries and 16 languages made the singular Gettier judgement at rates between 70% and 90%.[5] Machery et al. (2018a) report similar cross-cultural invariance: 86% of US respondents issued the singular Gettier judgement; 95% of Brazilians; 88% of Indians; 91% of Japanese.

Hospital represents an important class of Gettier cases. In the terminology of Turri (2019), it exhibits the structure: *no detect with replacement*.[6] Though the agent is reasonable to believe the proposition in question, they fail to genuinely detect its truth. The *presumed* truthmaker for the proposition has not in fact been realised; it is true in virtue of a 'replacement' truthmaker. Paul justifiably believes his wife is hospitalised, on the basis of a reasonable presumption that she was admitted to University. His presumption is incorrect: she was admitted to Metropolitan.

[4] The studies test both respondents' inclination to choose between a knowledge attribution and a straightforward ignorance attribution, and their inclination to choose between a knowledge attribution and describing the agent as merely having the impression that they know. The latter seems more revealing.

[5] Israeli Bedouins were an outlier; Machery et al. (2017) advise caution in light of a small sample size.

[6] These are 'apparent evidence' cases, in the terminology of Starmans & Friedman (2012).

This class is doubly notable. First, it plausibly includes the original counterexamples of Gettier (1963). Hence, the philosophical work achieved by Gettier's paper is equally achieved by the robust inducement of a singular Gettier judgement by Hospital. Second, there is evidence that cases in this class tend to induce the singular Gettier judgement with striking frequency: see Starmans & Friedman (2012); Turri (2013); Turri et al. (2015) for a selection.[7] This contrasts, Turri et al. (2015) show, with Gettier cases with so-called *detection with failed threat* structure (e.g. the fake-barn cases of Goldman (1976)) or *detection with replacement* structure (e.g. the 'authentic evidence' cases of Starmans & Friedman (2012)). Turri (2019) rightly cautions: that a certain type of Gettier case induces (or fails to induce) largely uniform judgement doesn't support conclusions about the abstract class of Gettier cases as a whole – in particular, those with very different epistemic structure. Let me stress that I nowhere assume that conclusions about Hospital translate into clear morals for, say, fake-barn cases (or *vice versa*).

Strikingly uniform folk judgement about Hospital doesn't indicate *accurate* judgement if folk epistemic judgement is systematically inaccurate. However, Williamson and Machery accept (what Alexander & Weinberg (2014) call) the *general reliability thesis*: blind-spots granted, folk epistemic judgement is generally accurate when evaluating suitably mundane cases.[8] Crucially, *non-exceptionalism* and the *general reliability thesis* yield:

> *Epistemic non-exceptionalism.* Absent specific defeat, a MoC judgement about a mundane case is rightly treated as expert judgement.

Epistemic non-exceptionalism would be questionable if promising accounts of MoC that entail it were elusive. Fortunately, as we saw in Chapter 9, Williamson (2007, ch. 6) offers such an account. To refresh our memory, the reasoning induced by Hospital is explicated roughly as:

W1 Hospital is (metaphysically) possible.

W2 If Hospital were the case, then someone would justifiably believe a true proposition without knowing it.

C1 Thus: it is (metaphysically) possible for someone to justifiably believe a true proposition without knowing it.

[7] Turri et al. (2015) observe a subtlety: the singular Gettier judgement seems notably suppressed if the actual ('replacement') truthmaker is suitably similar to the presumed truthmaker. The divergence dissipates for presentations that help respondents track underlying epistemic structure (Turri, 2013).

[8] Williamson (2018) argues that scepticism about 'philosophical intuition' is implausible if understood to encompass large swathes of mundane judgement (i.e., if one accepts non-exceptionalism). As for Machery: "[Radical restrictionism] is not a skepticism about judgment in general or, more narrowly, about the judgments concerning the topics of philosophical interest – e.g. knowledge, causation, permissibility, or personal identity" (Machery, 2017, p. 7).

C2 Thus: it is not (metaphysically) necessary that one knows p just in case p is true and one justifiably believes p.

Generally, Gettier-reasoning proceeds as follows: the subject judges both that the described case is possible (**W1**) and that if it were to occur, then someone would have a justified belief in true proposition p without knowledge of p (**W2**). The subject thereby draws a singular Gettier judgement (**C1**). The Gettier conclusion follows (**C2**).

W1 is justified by whatever justifies ordinary objective possibility claims (perhaps: reality-oriented imagination or ampliative reasoning). Williamson proposes that **W2** is justified via an exercise of reality-oriented imagination, guiding a simulated rational belief update: "one supposes the antecedent and develops the supposition, adding further judgments within the supposition by reasoning, offline predictive mechanisms, and other offline judgments" (2007, pp. 152-153). What grounds the accuracy of such simulations? For Gettier, we can partly appeal to our ordinary capacity for mindreading (Nagel, 2012). Indeed, given an *actual* Gettier case, the modal and counterfactual aspects of the reasoning are trivialised, with **W2**'s justification plausibly collapsing into mere mindreading.

Williamson's account has met resistance, but this needn't distracted us. First, it is 'proof-of-concept' for the Williamsonian approach, whatever refinements await. Second, the objections chiefly target the appeal to counterfactual reasoning, but such worries can be postponed by focusing on actualised Gettier cases. Third, the chief criticisms may not necessitate radical refinement. Remember, the account has been criticised for erroneously predicting that *deviant realisations* can defeat Gettier-reasoning (Ichikawa & Jarvis, 2009; Malmgren, 2011). A deviant realisation of Hospital satisfies its bare description but includes details that necessitate that the agent does *not* have a justified true belief without knowledge (e.g., Paul knows by an *unmentioned* source that his wife is in hospital). Now suppose that (only) deviant realisations are actual. Thus **W2** is false, and the Williamsonian must conclude that the Gettier-reasoning fails. This is counter-intuitive: if deviant realisations are actualised, this seems *irrelevant* to Gettier's force. Here are three strategies for amending Williamson's analysis.[9] The first targets the appeal to a counterfactual conditional, perhaps deploying a more subtle conditional (Geddes, 2017). The second amends the content of the counterfactual: perhaps the consequent is better explained as the stronger 'someone would justifiably believe a true proposition on grounds that are not sufficient for knowledge' (Sosa, 2017). The third questions whether Hospital is rightly taken as the input for the Gettier-reasoning: perhaps there is a gap between it and the intended extension thereof that the philosopher successfully communicates. Clearing this gap seems a job for a general theory of communication (see Chapter 9).

The account has advantages that refinements should arguably preserve. *Fit with*

[9] These were elaborately discussed in Chapter 9.

pre-theory: the account echoes a pre-theoretic description of participating in a Gettier thought experiment: the given text is a springboard for imagining a scenario that one judges to have certain epistemic features. *Non-exceptionalism:* understanding reality-oriented imagination as a form of simulation that bears on the epistemology of counterfactuals aligns with developments in cognitive science and psychology. Similar remarks apply to mindreading.[10] Moreover, counterfactual, possibility and epistemic judgements are ordinary phenomena with a plausible evolutionary purpose. *Possibility of success:* the argument from **W1** and **W2** to **C1** and **C2** is valid (on standard semantics). Further, general scepticism about such premises balloons into an implausible scepticism about everyday modal and counterfactual claims (see Williamson, 2016a,c). In particular, typical Gettier cases seemingly evoke mundane possibilities and everyday epistemic notions. *Possibility of defeat (i.e. fallibilism):* Since ordinary modal, counterfactual and epistemic judgements are fallible, Gettier-reasoning is predicted to be fallible. No appeal is made to infallible 'modal vision', 'rationalistic intuition', or 'raw conceptual competence' (e.g., Bealer, 1998; BonJour, 1998; Bealer, 2002; Sosa, 2007). This accommodates scepticism about applications of MoC where far-fetched possibilities are evoked or subjects lack requisite conceptual competence or background knowledge. (Compare Hospital to thought experiments that suspend the laws of nature or mention zombies.) Thus, *modal modesty* (or *moderate modal scepticism*) is accommodated, à la Van Inwagen (1998) and Hawke (2011, 2017). As Williamson puts it, "we are more reliable in evaluating some kinds [of counterfactuals] than others. [. . .] We may be correspondingly more reliable in evaluating the possibility of everyday scenarios than of 'far-out' ones, and extra caution may be called for in the latter case" (2007, p. 164).

Further alignment with Macherian commitments is now evident. Assuming that the experimental results collected in Machery (2017) indicate that epistemic peers are genuinely disagreeing when confronted with philosophical cases, a non-exceptionalist account of MoC must apparently accommodate blameless error, i.e., fallibilism. Further, Machery endorses moderate modal scepticism. Explicitly, Machery (2017, sec. 6.1.1) advocates scepticism towards (what he calls) *modally immodest philosophical theories*: theories committed to ambitious metaphysical necessities of peculiar philosophical interest. In support, Machery (2017, sec. 6.2) argues that stress-testing such theories requires an ability we lack: to reliably survey unusual, atypical, and remote possibilities. Thus, his advocacy of modal modesty is grounded in a moderate modal scepticism, which he in turn grounds in MoC's purported unreliability.

In this chapter, we will draw two main morals. First, the basic commitments of the Williamsonian F-line and Macherian X-line are largely complementary. (Section 10.3 exploits epistemic non-exceptionalism and the general reliability thesis;

[10] Gallese & Goldman (1998); Currie & Ravenscroft (2002); Nichols & Stich (2003); Goldman (2006); Gallese (2007). Also see Nagel (2012) and the extensive references therein.

Section 10.5 revisits modal modesty.) Second, assuming these commitments, the demographic data reported by Machery et al. (2017, 2018a,b) and the account of MoC in Williamson (2007) render it eminently plausible that Hospital-like Gettier cases induce reliable judgement.

10.3 Macherian Pessimism

At this point, it might be puzzling how a Macherian could be pessimistic about the reliability of Gettier. Two arguments for pessimism about MoC can be extracted from Machery (2017). In this section, I explicitly apply such arguments to Gettier, and respond.

Worrying data: Judgement in response to Gettier is significantly influenced by *mere* presentation: in particular, framing (Machery, 2017, ch. 2). Furthermore, particular presentations cannot be singled out as promoting accurate judgement. Thus, the Gettier judgement should be rejected as unreliable across the board.

Philosophy is disturbing: Relative to traditional philosophical aims, philosophically interesting cases generally have *disturbing characteristics* that promote unreliable judgement (Machery, 2017, ch. 3). Furthermore, Gettier is no exception: Gettier cases invariably have (at least) one of these characteristics. Thus, the Gettier judgement should be rejected as unreliable across the board.

In response to the first, I *conditionally* deny the second premise: it is reasonable (given epistemic non-exceptionalism) to take *certain* Gettier cases as evincing accurate judgement, *if* there aren't independent reasons to think Gettier cases are intrinsically disturbing. The second argument, I suggest, is thus the more basic of the two. In response to it, I again deny the second premise: Gettier cases don't characteristically exhibit any of the disturbing characteristics identified by Machery. In the remainder of this section, I will discuss the worrying data argument and in the next section I will turn to the disturbing characteristics argument.

10.3.1 Gettier and Framing

Does the worrying data cast doubt on the reliability of Gettier-reasoning? To focus the discussion, let's concentrate on the data issued by Study 2 of (Machery et al., 2018b).[11] Here, 85% of respondents judge that Paul in Hospital has the impression

[11] Other studies would serve equally well, though specific points might require tailoring. Machery et al. (2018b) offer evidence that Gettier is subject to order effects. Machery (2017, sec. 2.6.2) rightly points out that the effect is small. Machery et al. (2018a) observe that a certain Gettier case (the 'Trip case') induces a singular Gettier judgement at a markedly lower rate than Hospital,

that he knows, but doesn't know; while only 63% of respondents judge similarly for the agent in Clock, a second Gettier case. Clock is a variant on the classic case due to Bertrand Russell. (Basically: a stopped clock happens to read 4 o'clock on its face. At 4 o'clock, a hapless agent observes the clock face and thereby forms a belief about the time.) What to conclude?

It seems doubtful that the right conclusion is that Gettier cases evoke significantly non-uniform or unreliable judgement, for this requires an unmotivated inductive step. The class of Gettier cases is large, varying over possible epistemic structures and narrative details. Absent an argument that our sample (Hospital and Clock) is representative, nothing rules out, for instance, that the vast majority of Gettier cases induce the singular Gettier judgement at a rate akin to Hospital, with Clock an outlier.

The conclusion is in doubt even if one grants the sample is representative, for it isn't clear that the data exhibits a framing effect in the first place. A framing effect is exhibited by two cases when (i) there is a statistically significant difference in how subjects respond and (ii) the cases differ *only* in superficial narrative details: with respect to philosophically relevant structure, they are equivalent. Let's grant that Hospital and Clock both deserve the title 'Gettier case'. However, Starmans & Friedman (2012) and Turri et al. (2015) caution that Gettier cases vary significantly in underlying epistemic structure. Hospital and Clock exemplify this. In Hospital, the agent believes a proposition ('My wife is in hospital') on the basis of a presumed truthmaker (she was admitted to University) that differs substantially from the actual truthmaker (she was admitted to Metropolitan). Clock doesn't share this feature. Further, the nature of the defect in the agent's information source differs. In Hospital, the agent consults a device (a call to the hospital) that is (known to be) generally reliable with respect to the salient domain (admittance facts), but is, as a matter of (bad) luck, misleading in this one instance. In Clock, the agent consults a device (the stuck clock) that is (surprisingly) highly unreliable with respect to the salient domain (time facts), but is, as a matter of (good) luck, accurate in this one instance.

The conclusion is doubtful even if one grants the sample is representative *and* issues a framing effect. Machery (2017, p. 104) offers the following criterion for judging unreliability: "the judgments elicited by a given case are unreliable provided that they are influenced by at least a demographic variable or a presentation variable and provided that this influence is large [enough]". Note, however, that Machery (2017, sec. 3.3.1, p. 108) doesn't think it suffices that the influence count as 'large' in terms of standard benchmarks from psychology. To see why, first note with Machery

despite having a similar underlying epistemic structure. Again, the divergence is less marked than between Hospital and Clock. Turri et al. (2015) and Turri (2019) report studies indicating that Gettier cases with differing epistemic structure yield a singular Gettier judgement at different rates. Beebe & Shea (2013) report a study indicating that the moral valence of the agent's actions affects how subjects respond to a Gettier case.

(2017, p. 46) that we are concerned with cases where "the dependent variable is a percentage (e.g., the percentage of people agreeing that the character does not know the relevant proposition in the situation described by a Gettier case)." Machery (2017, pp. 45-47) deems the independent variable's effect size as 'large', relative to standard benchmarks, when the absolute difference between the percentages under two conditions exceeds 30%. Let's say, in this case, that the variable's influence is *significant*; assuring one that the observed effect doesn't merely reflect noisy data. (To illustrate: for Hospital and Clock, the difference in percentage is 22%, indicating only 'moderate' significance.)[12] However, 'significance' is then neither necessary nor sufficient for concluding that the population's judgement is unreliable. Consider sub-populations A and B, each making up 50% of the total population. If 100% of A-respondents and 70% of B-respondents answer 'yes' to polar question Q, then the influence of sub-population membership is significant, but, overall, 85% of the population answer 'yes'. If the correct answer is unknown, we can merely conclude that the population is *either* largely reliable on Q *or* largely unreliable. Further, if 52% of A-respondents and 48% of B-respondents answer 'yes', the difference in response is not significant, but the average response matches chance. The population is, on average, unreliable.

Thus, Machery (2017, sec. 3.3.1) proposes we attend to *average response*:[13] a variable has a large enough effect for determining unreliability when, in the aggregate (across different values of the variable), the distribution of responses is substantially mixed, i.e., the probability of any given response is sufficiently close to chance. That is, when the influence of the variable is accounted for, disagreement is stark.

To illustrate: suppose that half the population are political conservatives and half are political liberals. Suppose that 100% of conservatives answer 'no' to 'Is global warming real?', while 100% of liberals answer 'yes'. Thus, the distribution of 'yes/no' answers is 50/50. One concludes: the effect size of the (pernicious) variable of political affiliation is large enough to conclude unreliability, since it produces widespread disagreement in the aggregate. (Further, if we don't know which of 'yes' or 'no' is right, and we cannot assume that one sub-population has special competence on the issue, then we cannot identify which sub-population has accurate judgement, so cannot ignore the overall unreliability of the population's judgement.) Second example: suppose that 80% of conservatives answer 'yes' to 'Is global warming real?', while 100% of liberals answer 'yes'. Then the probability that a random member of the population will answer 'yes' is 90%: significant agreement

[12] Additionally, when one looks at standard normal curve test from statistical power analysis to measure the effect size of "the difference between two independent proportions" (Cohen, 1992, p. 157), which is gives the largest effect size of the discussed these methods, the effect size is still only medium (given $h_h = .85$ and $h_c = .63$, the arcsine transformation $\varphi = .51$) (see Cohen, 1988a, ch. 6 and Cohen, 1992, p. 157, Table 1). Thanks to Rob Schoonen for helpful discussion on this point.

[13] Relatedly, Machery responds to criticism from Demaree-Cotton (2016) – who argues that he concludes unreliability too quickly – that she "does not address the issue [of effect size] from the right angle" (2017, p. 108).

is exhibited. Hence, we shouldn't take the effect size as large enough (despite a 20% difference between groups) and shouldn't conclude that the population's aggregate judgement is unreliable.

Now compare Hospital and Clock. Here, the aggregate probability of a certain response is presumably calculated as the probability that a random member of the population gives that answer after being assigned Hospital or Clock with a coin flip.[14] If the experimental data is representative, the probability that 'mere impression of knowledge' is chosen over 'knowledge' is thus 74%. This represents notable agreement. (Machery presumably agrees: compare the 'room colour' example discussed by Machery, 2017, p. 104.) So why conclude significant unreliability, rather than lightly tempering one's credence that 'mere impression of knowledge' is the right answer?

Turn to my main argument, which is maximally concessive to Machery. Let's grant that the data indicates that Gettier-reasoning is significantly unreliable in the aggregate. Nevertheless, a question remains as to the exact conclusion this warrants.

Option 1: Judgement in response to Gettier cases is not terribly reliable in the aggregate.

Option 2: While judgement in response to Gettier cases is not terribly reliable in the aggregate, judgement relative to certain Gettier cases (or presentations thereof) is reliable.

Option 2 is a stronger hypothesis, and better explains the overall data. As noted previously, there is independent evidence that judgement induced by certain (presentations of) Gettier cases yields significant agreement across diverse demographics (Machery et al., 2018a,b). This uniformity is explained by Option 2 and left mysterious by Option 1. Certainly, if Gettier-reasoning were invariably unsystematic, then robust agreement on any particular Gettier case would be extremely surprising. So Option 2 should be accepted over Option 1, on abductive grounds.[15]

A fortiori, one shouldn't *suspend* judgement on the question of reliability (as a moderate restrictionist might advocate).[16] There is a good reason to take judgement induced by certain cases as reliable: this best explains a striking regularity.

But what of the possibility that significant agreement on a particular Gettier case indicates that our judgement is *systematically inaccurate* in that case? If this

[14] Let $\mathtt{Pr}(K|H)$ be the objective probability that a random respondent selects 'knowledge' on the condition they were assigned Hospital; let $\mathtt{Pr}(K|C)$ be the probability that they select 'knowledge' on the condition they were assigned Clock. Then the aggregate probability of the 'knowledge' response is $0.5 \times \mathtt{Pr}(K|H) + 0.5 \times \mathtt{Pr}(K|C)$, i.e. $0.5 \times 0.15 + 0.5 \times 0.37$, i.e. 0.26.

[15] Machery (2017, p. 106) claims that "it is hard to see which of the frames or which of the orders of presentation would make it more likely that people get it right about the situations described by philosophical cases". This is exactly what I deny.

[16] Notable moderate restrictionists happily concede that Gettier-reasoning can be reliable: see Alexander & Weinberg (2014) and Weinberg (2017).

were a serious possibility, then Option 3 could be deployed to explain the data, on a par with Option 2.

Option 3: Judgement in response to Gettier cases is not terribly reliable in the aggregate, and judgement relative to certain Gettier cases (or presentations thereof) is systematically inaccurate, generating an *epistemic illusion.*

However, an epistemic non-exceptionalist should *not* take Option 3 seriously without specific support for it over Option 2. Absent specific evidence that a certain (presentation of a) Gettier case corrupts judgement, she observes a basic confidence in ordinary judgement. If the case generates widespread agreement (relative to a large and diverse population of individuals), the presumption should be that ordinary judgement has here largely yielded accurate ('expert') judgement, as is typical for ordinary cases. Compare a toy example: suppose that half of the population of *climate scientists* are liberals, half are conservatives. It turns out that 98% of the former answer 'yes' to 'Is climate change real?', compared to only 60% of the latter. The uniformity among liberals is striking. Should we posit that their judgement is systematically inaccurate (wholly corrupted by political brainwashing)? This is excessively sceptical, in the absence of specific evidence. The normal presumptions stand until defeated: a scientist's judgement is normally expert, and expert judgement generally converges. Thus, the best explanation for the uniform liberal judgement is that it is accurate: the liberal experts judge exactly as we would expect experts to judge (striking consensus); while the conservative experts judge as we would expect experts to judge under the influence of disturbing factors (a mixed response).[17]

Is there independent reason to think that Gettier-reasoning typically exhibits peculiarities that jeopardise ordinary judgement? Were the answer 'yes', Option 3 would be live. I'll argue 'no' with respect to the 'disturbing characteristics' proposed by Machery (2017).

10.4 Is Gettier Disturbing?

Machery (2017, ch. 3.5) argues that philosophically interesting cases typically have one of three *disturbing characteristics* that promote unreliable judgement:

[17] Should we suspend judgement simply because there is significant disagreement between epistemic peers; indeed, presumed experts? (Cf. Machery, 2017, ch. 4.) I suggest we shouldn't. *Mere* disagreement between peers needn't prompt suspension of judgement: if 98% of experts agree on a question, one should accept the consensus without hesitation, despite some dissent. Generally, it is plausible that one should calibrate one's credence in line with the strength of consensus among experts. What's more, there are plausibly cases where one should give more weight to certain large sub-classes of expert, e.g. as in our toy example, when the sub-class of liberal experts judge as we expect experts should judge (i.e. with broad consensus), as opposed to the sub-class of conservative experts, who judge as we expect experts to judge in the presence of disturbing factors.

Entanglement: Judgement of the case is influenced by its superficial content. That is, arbitrary narrative details (that merely render the case concrete and vivid) influence our judgement, though they have no real bearing on the issue the case is intended to investigate.

Unusualness: The case describes an unusual situation, relative to the demands of ordinary life. Ordinary life doesn't offer opportunities to exercise judgement in such situations (not even unrealised opportunities), so we cannot assume ordinary judgement is primed for them.[18]

Atypicality: The case pulls apart properties that generally co-occur in ordinary life, sabotaging the heuristics of ordinary judgement and encouraging *ad hoc* responses.

It is explicable that philosophically interesting cases tend to have these features. Philosophy investigates phenomena that, while familiar and fundamental, puzzle us on close inspection. We engage in philosophical reflection precisely because we struggle to delineate core features. It is therefore difficult to guard against (or correct) entanglement. Further, philosophical theories often target necessary truths, with rival theories often agreeing on everyday cases. Such theories can only be stress-tested with unusual or atypical cases.

I'll discuss each disturbing characteristic in relation to Gettier, in turn.

10.4.1 Entanglement

Let's grant that philosophical cases face a *threat* of **Entanglement**: it is hard to rule out that any particular judgement is subject to entanglement. Further, I tentatively grant that there is specific evidence of entanglement in the case of Gettier: as noted, Machery et al. (2018b) report that responses to certain Gettier cases are influenced by merely presentational factors.[19]

Given the general reliability thesis, one must deny that the mere threat of entanglement casts doubt on the reliability of Gettier-reasoning. If it did, there would be similar grounds for doubting the reliability of countless ordinary epistemic judgements: the latter seem no less susceptible to entanglement. You see Sam reading the headline of today's New York Times. The headline states that Clinton lost the election. Sam is, in your experience, an affable and reasonable person. Further, you are aware of the Times' reputation for journalistic excellence and find it an enjoyable

[18] Cf. Weinberg (2017, p. 265): the relevant sense of 'unusualness' needn't "concern the frequency of the occurrence of Gettier-type situations, but the frequency of epistemic evaluations of Gettier-type situations, in which the relevant aspects of the situation are recognised and even capable of being brought into the evaluation."

[19] Though recall Section 10.3.1's reservations in concluding too hastily that two Gettier cases are equivalent with respect to core philosophical features.

read. You judge (rightly) that Sam thereby knows that Clinton lost the election. But the threat of entanglement is present. Absent general confidence in ordinary epistemic judgement, nothing rules out the possibility that one's judgement has here been influenced by epistemically irrelevant features of the situation (say, one's warm feelings for Sam or the New York Times). As usual, it is difficult to *exactly* delineate the features of the situation that make the knowledge ascription reasonable, so a more cautious assessment of Sam's epistemic state is elusive.

What of the specific evidence that presentation influences Gettier-reasoning? Remember the conclusion from Section 10.3.1: given epistemic non-exceptionalism, the best explanation of the *overall* data is that only *certain* Gettier cases (or presentations thereof) are likely entangled. This suggests that adverse presentation effects can be ameliorated by a judicious selection of presentational features (and that experimental philosophy provides useful tools for identifying them). Call those Gettier cases that elicit markedly stable judgement *sober*. Going forward, I focus on such and assume Hospital is among them.

10.4.2 Unusualness

That Gettier cases are unusual has initial support, as Weinberg (2017, sec. 3) notes. Anecdotally, philosophy students find them surprising on first encounter. Some need help to grasp their structure: rushing their introduction seems a pedagogical error. Experimentally, Turri (2013) reports that judgements about Gettier cases converge much more readily if their structure is presented with extra perspicuity. There is evidence, then, that Gettier cases don't regularly emerge for evaluation in ordinary life, and ordinary faculties aren't always primed to notice and properly assess them.

It doesn't follow that (Hospital-like) Gettier cases are intrinsically disturbing. To show this, we decompose Hospital.

Component 1: *Justified belief without 'sensitivity'.*[20]

> Starting to worry that something might have happened to his wife, Paul Jones decided to call some local hospitals to ask whether any patient by the name of "Mary Jones" had been admitted that evening. At the University Hospital, the person who answered his call confirmed that someone by that name had been admitted with major but not life-threatening injuries following a car crash. Paul grabbed his coat and rushed out to drive to University Hospital. As it turned out, the patient at University Hospital was not Paul's wife, but another woman with the same name.
>
> ***Judgement:*** *Paul didn't come to know anything about his wife via the call, but it led him to justifiably/reasonably/blamelessly believe she was hospitalised.*

[20] That is, with merely 'apparent evidence', in the terminology of Starmans & Friedman (2012), or 'without detection' in the terminology of Turri et al. (2015); Turri (2019).

Component 2: *True belief.*

Paul's wife had a heart attack as she was leaving work, and was actually receiving treatment in Metropolitan Hospital, a few miles away.

Judgement: *Paul had a true belief if he believed his wife was hospitalised.*

Component 1 yields, by itself, a key judgement: Paul's ignorance. Strikingly, the truth value of 'Mary Jones is in hospital' needn't be specified for this judgement to be apt. A tempting conclusion: the truth value is *irrelevant.* A ready explanation: misleading appearances aside, University's admission roster holds no information about Paul's wife, and sources that are uninformative about X don't induce knowledge about X.

The general phenomenon is familiar and mundane. Suppose Ann asks Bob, a trustworthy person: "Does Carol eat meat?" Bob sincerely replies: "No, Carol is vegetarian. She told me so". However, Ann and Bob are speaking at cross purposes: Ann is talking about *Carol Jones*; Bob about *Carol Smith.* Indeed, he doesn't know anything about (doesn't hold information concerning) the dietary preferences of Carol Jones. Ann might thereby reasonably believe Carol Jones is vegetarian, but this isn't knowledge; Bob didn't communicate any knowledge *about Carol Jones.* Whether or not Carol Jones is in fact vegetarian seems irrelevant to this mundane assessment. Another instance: Ann asks Bob: "Do all the conference speakers eat meat?" Bob sincerely replies: "No, one of them told me she is vegetarian". However, Bob is talking about Carol Smith: he mistakenly believes she is a conference speaker. Indeed, he doesn't know the dietary preferences of any conference speaker. Ann forms a reasonable belief that not every conference speaker eats meat. This isn't knowledge; Bob didn't have any to communicate. Whether any speaker is in fact vegetarian is irrelevant.

Further, Components 1 and 2 are, on their face, simple and mundane. Assuming the general reliability thesis, ordinary judgement is primed for such circumstances: absent defeating considerations, our assessment is trustworthy.

Of course, situations akin to Component 1 and 2 might occur infrequently. If so, they are unusual, in a straightforward sense. Does *this* defeat default confidence in our immediate judgements? No – it rather illustrates that low probability events can be mundane and, therefore, apt for reliable judgement. As Williamson (2016c, sec. 2.3) observes, to assume that low probability events invariably disturb ordinary judgement is markedly sceptical: just about any situation is of low probability under the right description. Indeed, it is evident that ordinary judgement *doesn't* collapse in the face of rare/unexpected events: if it did, we would be severely impeded in ordinary life.

If there is anything *notably* intriguing and unusual, it is the *combination* of Component 1 and 2. Mere combination can introduce two complications: lowered probability and heightened complexity. But, again, ordinary judgement isn't so brittle as to collapse in the face of lightly improbable combinations of ordinary situations. Sam reads in the New York Times that Clinton lost the election. Conclusion:

she knows Clinton lost. Blake reads in the New York Times that Clinton lost the election. Conclusion: she knows Clinton lost. Coincidentally, they read exactly the same copy of the NYT (at a certain doctor's waiting room; they both fell sick that day). We wouldn't and shouldn't retract our initial judgements of knowledge simply because of this coincidence. Rare combinations of mundane elements are sometimes mundane. Similar remarks apply to complexity introduction. We face complex situations in ordinary life (e.g. a busy city street). Navigating them requires skills in complexity management: selective attention and careful bookkeeping. An agent that lacks these is again severely impeded, certainly in high stakes situations. So, if complexity *invariably* disturbed ordinary judgement, the general reliability thesis would be undermined. Of course, Turri (2013) provides *prima facie* evidence that the complexity of some Gettier cases disturbs ordinary judgement. Unsurprisingly, this is ameliorated with a careful presentation (explaining why introducing Gettier cases to students requires care). Anyway, the experimental results indicate Hospital doesn't fall prey to such disturbance.

At any rate, even if the combination of Component 1 and 2 *could* lead to confusion, a simple strategy safeguards accuracy: be careful to judge the components individually and then conjoin the judgements. Could the combination of Component 1 and 2 somehow defeat the considerations that render the corresponding judgements individually apt? This strains credulity: the respective considerations seem decisive. Again, consider Component 1: it seems obvious that knowledge about X cannot accrue from a source that carries no information about X – no matter the circumstances of X.

So, is Hospital *disturbingly* unusual? This conclusion isn't licensed simply because it involves rare events or relative complexity. It seems a harmless combination of simple mundane elements: ordinary judgement is presumably expert here, a matter of merging individual judgements about Component 1 and 2. No experimental result defeats this presumption.

Weinberg (2017, sec. 3) proposes a more subtle reason to take (the simple elements of) Gettier cases as disturbingly unusual: they hinge on information about the 'specific inferential pathways' taken by the Gettierised agent. (He continues: "And it seems to me we only in the rarest of circumstances are in a situation to [furthermore] know that [the agent's] belief might be true, while also being aware of a range of possible truthmakers for that belief" (idem, p. 265).) Weinberg suggests there is a profound lack of such information in ordinary life. In one of the original cases of Gettier (1963), an agent uses disjunction-introduction to infer 'Jones owns a Ford or Brown is in Barcelona' from 'Jones owns a Ford', where 'Brown is in Barcelona' was randomly selected. It is hard to think of mundane situations where someone transparently reasons like this.

However, to claim mundane situations never yield information about 'specific inferential pathways', broadly understood, is to exaggerate. Ordinary speakers often report their reasoning for evaluation. Ann: "Someone in the office is vegetarian".

Dave: "How do you know?" Ann: "Bob is Carol's good friend and told me she is vegetarian". One judges: Ann believes that someone in the office is a vegetarian, on the basis (of her belief) that Carol is. One judges: she knows the former if she knows the latter, which hinges on whether Bob knew it. This is exceedingly mundane. (As is observing that Ann's belief that someone is vegetarian might be true, and could be made true by multiple possible situations.)

Grant that disjunction-introduction yields strange reasoning. Not all Gettier cases involve such strangeness. Another classic case from Gettier (1963) hinges, less artificially, on existential-introduction. Hospital induces a perfectly ordinary judgement about an agent's reasoning: Paul believes his wife was admitted to University, on the basis that her name is on the admission roster.

In short, Hospital might be unusual, but, assuming general reliability, we shouldn't take it as *disturbingly* unusual: unusual in any sense that undermines ordinary judgement. To generalise: absent specific defeat, cases in this structural family shouldn't be counted by an epistemic non-exceptionalist as disturbingly unusual if constructed from simple mundane elements, presented with perspicuity, and assessed with care.

10.4.3 Atypicality

Turn to **Atypicality**. Machery worries about situations where there is a package of features, e.g., a, b, c, that typically indicates Φ, and ordinary judgement exploits this as a mere *heuristic*. Thus one shouldn't conclude from our ordinary practice that any of a, b or c is *necessary* for the truth of Φ, nor that ordinary judgement is reliable when the package is pulled apart. Hence, philosophically interesting cases that fracture the package yield dubious judgements. In the case of Gettier, the typical package 'truth + justification + sensitive belief' indicates knowledge, and serves as a heuristic for ordinary judgement. (Let's grant these claims.) But Gettier pulls sensitivity (whatever it is) apart from truth and justification. Hence, the worry goes, Gettier induces unreliable judgement.

In response, two points: (i) splitting a typical package doesn't *necessarily* lead to unreliable judgement; (ii) Gettier plausibly investigates exactly this sort of split (i.e., where reliability is not undermined).[21] To see (i), consider: it is easy to think of ordinary situations where justification is present without truth. Here, a typical package is pulled apart. But we shouldn't conclude that judgement in these situations is unreliable, since lack of truth is an ordinary, decisive marker of ignorance (as Machery, 2017, sec. 3.6.3 notes). In support of (ii), I suggest that lack of sensitivity (whatever exactly it is) is analogous to lack of truth: an ordinary, decisive marker of ignorance. Again consider Component 1: a mundane situation where we judge an agent as ignorant, given a lack of sensitive belief. Despite being hard to

[21] See Williamson (2016c, sec. 2.3).

make precise, the rationale for this judgement is again easily gestured at. Though Paul is unaware of it, the admission roster issues misleading evidence concerning his wife. Indeed, misleading appearances aside, it clearly carries *no* information about his wife. Agents that form beliefs on the basis of a (relevantly) bereft information source don't thereby acquire knowledge. Compare: an agent that forms beliefs about a celebrity's lifestyle on the basis of The National Enquirer doesn't thereby accrue knowledge. The badness of the source is decisive: it doesn't (seem to) matter if the belief happens to be true or if the agent has somehow been convinced to consider the National Enquirer trustworthy.[22]

In short, Gettier cases like Hospital might be atypical, but, assuming general reliability, one shouldn't conclude a *disturbing* atypicality.

In sum: there seems to be no compelling reason for an epistemic non-exceptionalist to take disturbing characteristics as intrinsic to (or typical of) Gettier cases in Hospital's structural family.

10.5 Upshot for Machery's Master Arguments

PwPB defends a severe conclusion: philosophers should abandon the traditional method of cases. Machery reasons inductively, using an inductive step:

> If the judgments elicited by most of the philosophical cases that have been examined by experimental philosophers are unreliable, then the judgments elicited by most philosophical cases are plausibly unreliable. (2017, p. 102)

He offers three lines of support for this claim:

1. The tested cases are typical examples of philosophical cases: they "possess many of the properties many philosophical cases possess" (2017, p. 109).
2. "[The tested cases] are canonical. They are famous, and, consciously or unconsciously, they function as templates or paradigms when philosophers write novel cases" (2017, pp. 109-110).
3. Philosophically interesting cases typically possess the *disturbing characteristics* discussed above, so its members are generally relevantly similar to the cases that have been tested (2017, sec. 3.5).

[22] In the terminology of Machery (2017, sec. 3.6.3), my claim is that sensitivity is a 'central component' of the concept of knowledge, just as falsehood is a central component of our concept of ignorance.

On this basis, the tested cases are claimed to be *representative* of the class of philosophically interesting cases.[23]

Should we accept the inductive step? I proceed on the assumption that my previous arguments have been successful: Gettier cases needn't be taken to generally exhibit disturbing characteristics; Gettier-reasoning (applied to sober cases) induces reliable judgement; and non-exceptionalists needn't find this mysterious, as a Williamsonian analysis illustrates. In particular, I assume this for Gettier cases with the underlying epistemic structure of Hospital, including those of Gettier (1963). This puts pressure on the inductive step.[24] Gettier cases are clearly philosophically interesting. They aren't intrinsically disturbing. They are typical. They are (especially) canonical: few thought experiments (even limiting ourselves to Hospital's class) have been as influential or elicited as much consensus among philosophers. Certainly, it is rash to assume that cases that are controversial among philosophers (precisely because they plausibly disturb ordinary judgement) better represent the broad class of philosophically interesting cases.[25] In short, even if 1 and 2 are true, 3 and the inductive step shouldn't be casually accepted: what rules out that philosophically interesting cases are frequently akin to sober Gettier cases like Hospital?

Even if one grants Machery's inductive step (and that most tested cases induce unreliable judgements), one can resist his severe conclusion. For he requires another conditional: if most philosophically interesting cases induce unreliable judgement, then MoC should be abandoned. But Gettier, it seems, showcases a class of cases for which MoC proves effective, with significant philosophical benefits in tow (as its influence attests). This success should be preserved and emulated. The experimental results are a signal for caution and reform. MoC shouldn't be abandoned, but recognised as fallible and utilised with discipline (and experimental checks). Gettier (Hospital-like cases in particular) represents a paradigm towards which MoC can and should aspire.

Machery (2017, sec. 5.6) is sceptical about the prospects for reforming MoC. Further, he anticipates objections to his inductive argument. He writes:

> Nor is it an objection that some philosophical cases may not possess any disturbing property. The claim is not that every philosophical case elicits a cognitive artifact or diverse responses, but that the kind of case

[23] Machery (2017, pp. 128-129) deploys a similar inductive step with similar justification in support of another argument for abandoning MoC: an argument from peer disagreement. Thus my critical remarks transfer to his second argument.

[24] Levin (2019) also suggests that Machery's inductive argument might be too quick.

[25] For instance, the zebra case of Dretske (1970); the 'fake barn' case of Goldman (1976); the zombie case of Chalmers (1996); et cetera. For clearly articulated suspicions about the force of some philosophers' knee-jerk judgements about these cases, see Van Inwagen (1998); Hawthorne (2004); Gendler & Hawthorne (2005).

> philosophers use for dialectical purpose tends, non-accidentally, to elicit cognitive artifacts or a diversity of responses. (2017, p. 183)

My own objections don't rest merely on the possible existence of philosophical cases that aren't disturbing: I am *not* fallaciously proposing that a single counterexample undermines a statistical or generic claim. My key claim is that certain *typical* and *canonical* philosophical cases don't possess disturbing properties. In this connection, let me emphasise that Machery doesn't deploy vanilla statistical-inductive reasoning: he doesn't base his conclusion that most philosophical cases elicit unreliable judgements on a (demonstrably) *random* and *suitably large* sample of tested philosophical cases (or, indeed, of tested typical and canonical cases). Nor does he establish the *relative degree* of typicality or canonicity for various philosophical cases, as would be essential for evaluating the plausible hypothesis that Hospital-like Gettier cases typify an especially large bulk of philosophical cases. Thus, he hasn't established that his sample of tested cases warrants generalisation to most or all philosophical cases; nor that Gettier isn't by itself a significant success story for MoC.

Machery continues:

> Nor is it compelling to respond that the advice to suspend judgment remains inapplicable until there is clear-cut evidence about what cases exactly are impugned by experimental-philosophy studies. First, we have provided reasons to believe that disturbing cases prime unreliability and disagreement. Second, even if we were unsure about how broadly to suspend judgment, we should still suspend judgment in response to all the cases in contemporary philosophy (except those known to be immune to demographic and presentation effects) because the cases examined by philosophers are typical and canonical. Similarly, if we find that some eggs are contaminated with Salmonella, we would stop eating eggs sold by the brand selling them, even if it is unclear whether all eggs are contaminated. (2017, p. 183)

In issuing a blanket ban on new applications of MoC (though he grants the possibility of cases that are immune to serious demographic and presentation effects), Machery underestimates our ability to (reasonably, defeasibly) discriminate between philosophical cases that are likely or unlikely to induce reliable judgement. Compare the debate induced by the proposal in Weinberg (2007) that epistemic judgement about philosophical cases isn't sufficiently *hopeful*: we lack robust error-detection mechanisms for regulating it. Ironically, Machery (2017, ch. 3) convincingly defuses generic worries about hopefulness. Further, studies reported by Wright (2010, 2013) suggest that ordinary respondents reliably register the presence of instability/unreliability in their epistemic judgements.[26] Machery has himself identified a rough but promising list of features that problematic cases typically exhibit: namely, the disturbing

[26] Machery (2017, p. 122) observes that the findings in Wright (2013) show that low confi-

characteristics (entanglement, unusualness, atypicality). If such characteristics are lacking (as far as one can tell), an epistemic non-exceptionalist assumes that ordinary judgement is primed to rule accurately on what appears to be an ordinary case. This assumption can, of course, be defeated by experimental investigation. Granted, some disturbing characteristics may be hard to discern: entanglement, for instance. However, our reservation in taking the mere threat of entanglement too seriously (Section 10.4.1) is again pertinent. Other characteristics seem easier to spot: modally exotic cases involving philosophical zombies or evil demons seem easily distinguished from relatively mundane cases like Hospital.

So the Salmonella analogy is inapt. Contrast a second case of egg contamination. In the summer of 2017, The Netherlands experienced a large scale contamination of eggs with *fipronil*, a poisonous insecticide (NOS, 2017a,b). The level of fipronil was so high in certain clusters of eggs that those eggs were inedible. But the National Health Organisation merely advised people to 'proceed with caution' when consuming eggs, rather than halt consumption altogether. This was sensible: it was reasonably clear which eggs were contaminated. Indeed, a serial number is printed on every egg, and the Dutch National Health Organisation was able to release a list of numbers for eggs that were reasonably suspected to be infected.

The same advice applies to MoC: a non-exceptionalist should proceed with caution, but to discontinue MoC entirely is an overreaction to the data.

10.5.1 Inductive Modal Modesty

Macherian pessimism about Gettier should be unconvincing to both Williamsonians and Macherians, in virtue of common ground: epistemic non-exceptionalism. The Williamsonian F-line gets a better handle on Gettier: suitably mundane cases like Hospital deploy ordinary possibility, counterfactual and epistemic judgement in the production of substantive philosophical knowledge. So much for Machery's (2017) claim that (traditional) MoC has not or cannot yield substantive philosophical conclusions, and should be shelved.

However, Machery (2017, Introduction) describes his critique of MoC as a detour on the way to his *main* conclusion that "resolving many traditional and contemporary philosophical issues is beyond our epistemic reach" (p.1); in particular, "modally immodest issues cannot be resolved, and modally immodest philosophical views [cannot be] supported" (p. 3). Philosophers, he worries, often pursue theories of knowledge, mind, personal identity, right action and free will that target ostentatious claims of metaphysical necessity. Machery (2017, sec. 6.1.1) offers this argument:

dence predicts unreliable/unstable judgement, but don't establish that high confidence predicts reliable/stable judgement. Thus, low confidence plausibly defeats the presumption that a judgement is stable/reliable. Our default trust in our (confident) judgement, meanwhile, can rest, for a Macherian, on the general reliability thesis.

M1. Many central philosophical issues are about metaphysical necessities, and resolving these issues requires establishing these necessities.

M2. Philosophers must appeal to unusual and atypical philosophical cases to establish these metaphysical necessities.

M3. We should suspend judgement about the situations described by current philosophical cases and, more generally, by unusual and atypical philosophical cases.

M4. There is no other way of learning about the pertinent metaphysical necessities and possibilities.

MC. Hence, there are many philosophical issues that we cannot resolve.

The arguments from this chapter allow us to reject **M3**: Hospital counts as a 'current philosophical case' that is, broadly speaking, unusual (Section 10.4.2) and atypical (Section 10.4.3), yet apt for judgement. Since Hospital represents a canonical and typical class of cases, **M3** shouldn't even be accepted *generically*.

However, a nearby argument is harder to dismiss. Gettier-reasoning is typically mundane and well-supported by empirical studies. This cannot be said for a large swathe of tested philosophical cases: *Truetemp, Switch, Transplant, Society of music lovers*, et cetera. Unlike Hospital, these don't seem to be pre-theoretically as (unlucky but) mundane: they are unusual or atypical in a plausibly disturbing sense. Indeed, empirical investigation reveals serious demographic and presentation effects (Machery, 2017, ch. 2). Suppose these cases are indeed canonical and typical examples of a larger class of *exotic* philosophical cases (in contrast to *mundane* philosophical cases). Indeed, they largely belong to a salient sub-class: modally *remote* cases, instantiated only in suitably 'distant' possible worlds. One may then deploy an inductive argument (analogous to but more modest than that in Section 10.5): MoC applied to *exotic* philosophical cases is unreliable. This supports:

M3*. We should suspend judgement about the situations described by exotic (e.g. remote) philosophical cases.

Here is a variant of **M2**:

M2*. Philosophers must appeal to exotic (e.g. remote) philosophical cases to establish these metaphysical necessities.

Replacing **M2** with **M2*** and **M3** with **M3*** yields a Macherian argument for **MC** that is untouched by the foregoing critique of this chapter. The tentative neo-Macherian moral: philosophers ought not abandon (substantive uses of) MoC, but limit it to (putatively) mundane cases that cannot support especially ambitious, modally immodest metaphysical theses. However, as Machery points out, philosophical "theories are often modally immodest: Their claims are often not primarily about how things actually are or about how things must be in worlds that obey the laws of nature; rather, they are often about how things must be, period" (2017, p. 186).

So, a version of Machery's core argument for rejecting modally immodest philosophy survives the critique of the preceding chapter. However, what exactly does it mean that philosophers (or ordinary humans for that matter) should be *modally modest*? Even though modal modesty, or 'moderate modal scepticism', is often appealed to (e.g., Van Inwagen, 1998; Williamson, 2007; Hawke, 2011; Fischer, 2016a; Leon, 2017; Machery, 2017), it is not always clear what philosophers have in mind when they appeal to it nor whether everyone has the same thing in mind. In particular, the modal modesty that Machery argues for seems subtly different from more common forms of modal modesty. We turn to these issues in the next, concluding chapter.

Conclusion

Chapter 11

Conclusion and Further Work: Modal Modesty

In general, there is a degree of doubt, and caution, and modesty, which, in all kinds of scrutiny and decision, ought for ever to accompany a just reasoner

– Hume, *Enquiry* (XII, III, p. 111)

The previous chapter concluded that non-exceptionalist philosophers do not need to shelve the use of thought experiments in its entirety. However, we also saw that a moderate version of the experimental philosophers' argument against thought experiments remained. We can come to justifiably believe some philosophically significant possibility claims, suitably mundane ones, yet this ability might be limited. That is, we have to be *modally modest.*

Modal modesty has come up throughout this dissertation. In this chapter I will first summarise the findings of this dissertation and then elaborate on modal modesty, the common thread of this dissertation. I will clarify what modal modesty could be and point to some future work with relation to it, highlighting a variety of potential interpretations, motivations, and consequences of modal modesty.

11.1 Conclusions of this Dissertation

In this dissertation, I set out to advance the debate on how we ordinarily have justified beliefs about what is possible. This work was divided into three part: Part I analysed imagination-based approaches to the epistemology of possibility; Part II examined similarity-based approaches; and Part III evaluated whether a cognitively plausible epistemology of possibility could ground a non-exceptionalist epistemology of thought experiments. Here, I briefly summarise the main findings.

In Part I, I critically evaluated three of the main theories of imagination for their potential to feature in imagination-based epistemologies of possibility. Chapters 3 and 4 argued that two of these – QALC imagination and pretense imagination, respectively – rely on forms of prior modal knowledge. This, I argued, results in these theories being unable to (ultimately) explain our knowledge of non-actual possibilities. In Chapter 5, I discussed appearance-based theories of imagination, which hold that imagination is the simulation of perceptual states. These theories do not rely on prior modal knowledge, but, I argued, require substantial assumptions to get beyond mere *phenomenal* evidence and provide justification for beliefs in the *objective* possibilities we are interested in. I presented a novel theory, which I called *embodied imagination*, where we take imagination to be *sensori-motor simulation*. I showed that this theory does not succumb to the issues raised against the appearance-based theories and can provide us with knowledge of objective, non-actual possibilities.

In Part II, I turned to the more recent *similarity-based* epistemologies of possibility. In Chapter 7, I examined the crucial notion of these theories, *relevant similarity*, in detail. I argued that one of the most prominent interpretations of relevant similarity, *predictive analogy*, leads similarity-based epistemologists a significant crossroads: either (i) they have to accept that the significant work is delegated to the epistemology of causation, with all the consequences of the particular theory of causation one accepts, or (ii) they need to develop an alternative to predictive analogy as a plausible ground for similarity reasoning. I opt for (ii) and in Chapter 8 propose my own theory of a similarity-based epistemology of possibility that is based on the notion of *kind*. I develop a particular technical notion of kindhood to support this inference and I discuss the epistemology of categorisation. I rely on findings from the psychology of reasoning to suggest that humans reason in accordance with a placeholder heuristic. Humans reason *as if* there is a core of properties that causes many of the other properties and behaviours shared by members of a kind, without needing to know the *explicit* causal relations involved.

In Part III, I analyse the use of hypothetical reasoning in philosophy, philosophical possibilities, and their role in philosophical thought experiments. In Chapter 9 I discussed Williamson's (2007) epistemology of thought experiments and one of the most prominent objections against it: the problem of deviant defeat. I proposed a solution that suggests that the analysis of thought experiments should take the intended enriched input as their starting point. This vindicates Williamson's original account, but also allows us to analyse the reasoning in thought experiments in terms of the Ramseyan indicative. In Chapter 10, I examined Machery's sceptical argument that we should *abandon* the use of thought experiments altogether. I argued that his conclusion is too radical: there are suitably mundane, philosophically interesting thought experiments that we can justifiably believe to be possible. This, I argued, has serious consequences for Machery's pessimistic inductive argument against the use of thought experiments in philosophy.

11.1.1 A Common Theme: Modal Modesty

In this chapter, I will elaborate on a common theme of this dissertation: *modal modesty*. So let me note some points where modal modesty has come up so far.

The two main positive proposals of this dissertation – embodied imagination and kind-based similarity reasoning – are naturally limited when it comes to the range of possibility claims they can justify. The nature of embodied imagination is such that when we imagine situations that are *within* our 'natural domain', imagination is taken to be reliable. Imagining possible actions, the consequences thereof, and anticipating the 'behaviour' of our environment in mundane situations are examples of such reliable imaginings. However, it is an open question – that to some extent depends on future findings of (embodied) cognitive science – how reliable imagination will be beyond these ordinary, mundane situations. Similarly, our ability to classify objects into kinds, and reason inductively on the basis of this, is very reliable when it concerns ordinary, mundane objects and situations – e.g., if I see one of my cats jump on the table, I can reliably conclude that my other cat *could* do so as well. However, it is not obvious that we can reliably apply this sort of reasoning to, or even correctly categorise objects involved in, exotic cases involving, e.g., disembodied ghouls or philosophical zombies.

The limits of our cognitively plausible epistemology of possibility have an effect on philosophy itself. Even though I concluded in Chapter 10 that (*pace* Machery) we *can* have justified beliefs about philosophically interesting thought experiments, a moderate version of Machery's argument for modally immodest philosophy survives. Philosophers ought not abandon (substantive uses of) thought experiments, but limit it to (putatively) mundane cases, which might not be able support especially ambitious, modally immodest metaphysical theses. Interestingly, the modal modesty that Machery (2017, ch. 6) argues for is subtly different from the kind of modal modesty resulting from embodied imagination and kinds-based similarity reasoning (the difference will be discussed in Section 11.3).

Modal modesty in and of itself hasn't received much attention in the literature yet (some exceptions are Hawke, 2017; Leon, 2017; and Strohminger & Yli-Vakkuri, 2018b). I think that if we want to properly develop a cognitively plausible epistemology of possibility, we need to take modal modesty seriously, develop it properly, and consider the consequences it might have on our overall approach to the epistemology of modality. In the remainder of this conclusion, I will put forth some thoughts for such future work. I will discuss some considerations on getting clear on what modal modesty might be; note a number of interesting varieties of modal modesty; and highlight different motivations for modal modesty. Finally, I conclude by noting some potential consequences of accepting modal modesty. All of these are exploratory remarks that deserve to be properly developed as we continue our search for understanding how it is that we know what is possible.

11.2 Getting Clear on Modal Modesty

Modal modesty, in general, concerns the *range* of our modal knowledge. The reconstructed argument from Machery, discussed in the previous chapter, gives us reason to believe that some form of modal modesty must be true. A different way to get to the same conclusion is taking the empirical data concerning more exotic cases from Machery (2017, ch. 2) at face value, in combination with our starting assumption that *radical* modal scepticism is false (see Chapter 1, Section 1.2). This suggests that *moderate* modal scepticism should be adopted (see, e.g., Hawke, 2017, sec. 11; Leon, 2017, sec. 2; and Machery, 2017, p. 188 for discussions of 'grades' of modal modesty/scepticism).

In this section, I will spell out what I take modal modesty to be. In order to do so, I will briefly say something on kinds of 'modesty' that are *not* what I intend.

11.2.1 The 'Modesty' of 'Modal Modesty'

First of all, I am *not* interested in a kind of *general* modesty – i.e., related to what it is to be a modest person in general (Driver, 1999). Rather, we are concerned with an instance of *epistemic modesty.*[1]

Talk of epistemic modesty might remind some of the literature concerning *peer disagreement*. This interpretation of epistemic modesty has it that "disagreement of others who have assessed the same evidence differently provides at least some [...] reason to be less confident in the conclusion we initially came to" (Christensen, 2013, p. 77). This is also not the kind of modesty that I am interested in. In particular, the epistemic modesty that some in the literature on disagreement recommend concerns the *lowering of one's credence* in certain propositions, whereas the kind of modal modesty that I am interested in focuses more on the *range* of our modal beliefs which we can be said to justifiably believe.[2]

[1] Though Driver (1999, p. 830) does talk about 'epistemic modesty', for her, this is an epistemic *analysis* of general modesty. Whereas we are concerned with modesty with regards to what one knows or can know.

[2] Though, there seem to be some similarities between this kind of epistemic modesty and the

Finally, there are those who take modal modesty, in particular when phrased in terms of 'modal scepticism', to concern the *meaningfulness* of modal notions or "about the reality of modal facts" (Machery, 2017, p. 188). This is again not the kind of modal modesty that I am interested in. The assumption, from Chapter 1 (Section 1.1), that modality is something mind-independent, can be taken as granting (even if just for the sake of the argument) that modal expressions are meaningful and that there is something that makes them true (or false). The modal modesty under discussion here concerns doubts about our knowledge of or our ability to know modal claims.

11.2.2 Modal Modesty

The modal modesty that I am interested in is something akin to the epistemic modesty in ethics discussed by Laskowski (2018). Laskowski points out that "[m]any prominent ethicists, [. . .], accept a kind of epistemic modesty thesis concerning our capacity to carry out the project of ethical theorizing" (idem, p. 1577). After rejecting the analogue of radical modal scepticism, Laskowski concludes with the following definition of epistemic modesty in ethics:

Modest Necessarily, for any subject S *like us in a world like ours* and any comprehensive ethical theory P, S does not know P.
(2018, p. 1588, original emphasis)

The idea that Laskowski aims to captures is that even though we might be able to theorise about some ethical situations, it is unclear whether we could theorise about *all* ethical situations, in particular about those that involve hypothetical situations far removed from our ordinary experience.

I take *modal modesty* to be something akin to Laskowski's *Modest*-thesis. In particular, I will focus on the kind of modal modesty discussed by Hawke (2011, 2017); Leon (2017); and Strohminger & Yli-Vakkuri (2018b). The canonical source for this kind of modal modesty is the seminal work of Van Inwagen's.

> My own view is that we often do know modal propositions, ones that are of use to us in everyday life and in science and even in philosophy, but do not and cannot know (at least by the exercise of our own unaided powers) modal propositions [. . .]. I have called this position 'modal skepticism.' This name was perhaps ill-chosen, since, as I have said, I think that we do know a lot of modal propositions, and in these post-Cartesian days, 'skeptic' suggests someone who contends that we know nothing or

kind of modal modesty that I am interested in (for example, in both cases might result in that we should be less certain of some of our philosophical views). Developing these similarities properly is left for future work.

> almost nothing. It should be remembered, however, that there has been another sort of skeptic: someone who contends that the world contains a great deal of institutionalized pretense to knowledge of remote matters concerning which knowledge is in fact not possible. [...] It is in this sense of the word that I am a modal 'skeptic.'
>
> (1998, p. 69, footnotes removed)

Van Inwagen here points out that the label 'modal scepticism' (later quantified to 'moderate modal scepticism') was "ill-chosen" and many have noted this since. For this reason, I decide to use the term 'modal modesty' for this position (remember footnote 11 in Chapter 1).[3] Leon nicely captures the *general* kind of 'moderateness' that is involved in modal modesty:

> [T]here is another sort of skepticism that doesn't write off the relevant class of beliefs due to general worries about its source or basis – the source or basis in question may well be capable of yielding knowledge or justified belief. The problem is that the source's capacity to justify beliefs is *severely limited*; in fact, its justification-conferring ability is limited to beliefs involving the practical concerns of daily life.
>
> (2017, p. 249, original emphasis)

Applying this to the epistemology of possibility, we get modal modesty. Hawke aptly describes a modally modest epistemologist as "one who holds that [...] while we have a great deal of basic, ordinary modal knowledge, our ability to establish more exotic possibility (or necessity) claims is importantly limited" (2017, p. 282). This kind of modal modesty often seems to be implicitly endorsed, yet is rarely explicitly discussed.[4]

Let me stress two things. First of all, modal modesty concerns the *range* or *scope* of our modal knowledge. We do have knowledge of or justified beliefs in possibility claims; it is just that the possibility claims of which we do, might not reach until the outskirts of modal space.[5]

Secondly, it is an interesting question in what sense we *cannot* come to know or resolve exotic possibilities. The usage of phrases such as 'subjects like us' and 'worlds like ours' suggest that this might not be unrestricted metaphysical necessity. Strohminger and Yli-Vakkuri point out that modal scepticism often focuses on

[3] The label is adapted from Machery (2017).

[4] Strohminger & Yli-Vakkuri (2018b) provide an excellent overview of Van Inwagen's modal modesty and point out that something like modal modesty is (somewhat) implicit in Williamson's (2007) epistemology of modality.

[5] Strohminger & Yli-Vakkuri (2018b) give their definition of modal modesty in similar terms. "An attitude of epistemic humility, however, seems to us at least as warranted in the epistemology of modality, when it comes to the knowability of the possibility of states of affairs that are distant from actuality" (idem, p. 319).

knowability *given our actual human cognitive capacities.* So, they suggest to focus on *human possibility* when we are considering modal modesty, where human possibility is a possibility that is "compossible with the cognitive (and other) capacities of human beings as they actually are" (Strohminger & Yli-Vakkuri, 2018b, p. 303). This restriction to 'human possibility' is in line with what authors such as Hawke (2017) and Leon (2017) seem to have in mind. Both point out that "our ability" to establish or justify beliefs in (exotic) possibilities is "limited" (Hawke, 2017, p. 282 and Leon, 2017, p. 249).

11.3 Varieties of Modal Modesty

At this point, it is important to stress that the kind of modal modesty described above (i.e., à la Van Inwagen, 1998; Hawke, 2011) is subtly different, in two interesting ways, from the modal modesty that Machery (2017) argues for. In this section, I will delineate a variety of forms of modal modesty that one might adopt.

11.3.1 Necessity Modesty

Machery argues that philosophical theories are immodest in that "[t]heir claims are often [...] *about how things must be*, period" (2017, p. 186, emphasis added). Machery seems to worry about philosophers' ability to make substantial claims about *necessities* rather than possibilities. This is also reflected in **M2/M2*** of his main argument (2017, sec. 6.1.1), where he talks about 'establishing metaphysical necessities'. Call this *Necessity Modesty.*

We can explain our knowledge of some necessities rather easily. We can explain how it is that I know that bachelors are necessarily unmarried men and that, necessarily, there is no barber who shaves all and only those men who don't shave themselves (Van Inwagen, 1998; Leon, 2017). Additionally, it seems uncontroversial that we have some basic mathematical knowledge (e.g., $2 + 2 = 4$), which plausibly counts as knowledge of necessities (even if Benacerraf (1973) worries are present when giving a full philosophical account of mathematical knowledge). However, it is much less obvious how I 'establish' or know what knowledge necessarily is (e.g., a particular kind of justified true belief) or that water is necessarily composed out of H_2O. So, modesty with regards to our knowledge of metaphysical necessities seems *prima facie* justified, is interesting, and deserves careful development and scrutiny.[6]

[6] Though it seems fair to worry about how philosophers establish or know of metaphysical necessities, given the focus on the epistemology of *possibility* of this dissertation, this worry is somewhat orthogonal to the discussions of this dissertation.

11.3.2 Situation Modesty

The modal modesty described in the previous section and developed by, e.g., Van Inwagen (1998); Hawke (2011, 2017); and Leon (2017) concerns our ability to judge certain *situations* as possible or not (this is the kind of modesty that the novel epistemologies of possibility of this dissertation most straightforwardly relate to). For example, we are able to judge (or know) that my coffee cup could be filled with coffee rather than being empty, yet it is unclear whether we can judge (know) that there could be disembodied ghouls or unconscious physical duplicates of us.[7] Call it *Situation Modesty*, as it involves our (lack of) knowledge that particular situations are possible.

11.3.3 Judgement Modesty

Interestingly, there is another kind of modal modesty implicit in Machery's work that concerns our knowledge of possibility and that is subtly different from situation modesty. Macherian modal modesty seems to focus less on our ability to know that a hypothetical situation is possible and more on our ability to make reliable, ordinary judgements *about* those hypothetical situations. For lack of a better word, call this *Judgement Modesty*. The Williamsonian account of thought experiments (see Chapter 9) helps to draw the distinction with situation modesty. Remember the reconstruction of our reasoning about philosophical thought experiments (Gettier cases discussed here, see page 206, repeated below):

W1 $\Diamond\exists x GC(x,p)$

W2 $\exists x GC(x,p) \mathrel{\Box\!\!\rightarrow} \exists x(JTB(x,p) \land \neg K(x,p))$

C1 $\Diamond\exists x(JTB(x,p) \land \neg K(x,p))$

Situation modesty worries about our ability to establish the possibility of certain remote hypothetical situations: premise **W1** in the reconstructed MoC reasoning is questioned. Judgement modesty, on the other hand, worries about the accuracy of our ordinary judgements about remote cases: **W2** is questioned. Thus, suspicion is raised about judgements about knowledge, right action or free will in response to *clearly possible but remote* cases. Another way to highlight the difference is through counterpossible conditionals (i.e., counterfactual conditionals with impossible antecedents).[8] If the impossible antecedent is both remote and only supports *some* consequents, the situation modesty worries about whether and when we can rightly *identify* a counterpossible conditional as such, whereas judgement modesty worries about whether and when we can rightly assess its *truth*.

[7] Strohminger & Yli-Vakkuri (2018b) make an interesting distinction between 'knowing whether p is possible' and 'knowing that p is possible'. I focus on the latter.

[8] See, for example, Nolan (1997) and Berto et al. (2018).

11.4 Motivations for Modal Modesty

Throughout the literature, though often implicit, one5 finds different motivations for accepting a form of modal modesty. In this section, I will briefly mention some of these. Future work should develop these motivations more fully, relate them properly to the different variations in modal modesty, and critically evaluate each of them.

11.4.1 The Disagreement Motivation

Borrowing from the literature on 'general' disagreement, one might think that a way to motivate modal modesty is by appealing to the (widespread) disagreement amongst people about the correct judgements in certain hypothetical situations. One might appeal to some of the empirical data from the experimental philosophy literature to support this motivation (Machery, 2017, ch. 2).

Even though one might use the appeal to disagreement about such cases to motivate *lowering one's credence* in such modal beliefs (Christensen, 2013), this does not rule out that, in principle, we could come to know the right judgement (i.e., that it is or is not possible).[9] An initial hypothesis for further work is that disagreement, by itself, does not motivate the stronger modal modesty where one takes ordinary agents not to be in a position to know certain exotic possibilities.

11.4.2 The Induction Motivation

Laskowski argues for our inability to (come to) know complete ethical theories by *pessimistic induction*. He points out that despite prolific ethical theorising, there is no consensus on what the "ethically significant features of the world" are (2018, p. 1587). This, he concludes, "constitutes strong inductive evidence for the claim that there will always be ethically significant features of the world of which we are unaware, which [in turn] suggests that we'll never [. . .] believe and hence know the true ethical theory" (ibid.). Machery (2017), as we saw in the previous chapter, aims to provide a similar pessimistic inductive argument against our reliability when it comes to making judgements about hypothetical philosophical situations. The rough idea of such a motivation for situation/judgement modesty is that in many cases we are mistaken in judging a situation to be possible/in the conclusions we draw from hypothetical situations. From this, one then inductively concludes that we cannot know whether a situation is possible or what conclusions to draw from hypothetical situations.

As discussed in Chapter 10 (Section 10.5.1), a straightforward version of such

[9] This suggests another variety of modal modesty: *credence modesty*. On such a theory of modal modesty, one would simply have lower credences in exotic possibilities than in mundane possibilities.

a pessimistic inductive argument does not survive the criticisms discussed in that chapter. However, a version supporting a *modest* conclusion does survive those criticisms and is supported by the empirical data from experimental philosophers.

11.4.3 The Ordinary Cognition Motivation

The most common motivation for accepting a form of modal modesty comes from the aim to provide a cognitively plausible epistemology of modality or from a prior acceptance of (modal) empiricism.[10] The reasoning here is as follows: our ordinary cognitive capacities and our empirical knowledge is mostly (reliable when) concerned with ordinary, mundane possibility claims, yet is it not obvious that we can rely on these things equally well when assessing more exotic or remote possibilities.[11]

Many epistemologists of possibility of an empiricist leaning hold something like this. For example, when talking about the "foreign terrain of absolute modality," Nichols points out that "the psychological systems are being used outside their natural domain," which means, he continues, that "there's less reason to think that they will be successful guides" (2006a, p. 253). Williamson (2007) holds something similar and is followed by Strohminger and Yli-Vakkuri.[12]

> [I]t seems plausible to us—although Williamson does not say this—that, at least generically speaking, the more distant a state of affairs p is from actuality, the more difficult it will be to imagine how things would be if p were to obtain in the amount of detail required for knowing that p does not counterfactually imply a contradiction.
>
> (Strohminger & Yli-Vakkuri, 2018b, p. 317)

Similar considerations lead Van Inwagen (1998); Hawke (2011); Strohminger (2015); Fischer (2016a); Hawke (2017); Leon (2017); and Roca-Royes (2017) to accept forms of modal modesty. (Though note that in the case of Van Inwagen (1998) and Fischer (2016a), this is because they only focus on *imagination*-based approaches, whereas for others it might be more generally due to the aims of cognitive plausibility.)

[10] It is an interesting question whether there is room for (or even a need for) modal modesty in more rationalist approaches. I will leave this aside for the purposes of this dissertation.

[11] This motivation subsumes arguments such as Van Inwagen's (1998) 'distance-analogy' and Fischer's (2016a) 'argument from epistemologies of possibility'. The former suggests that just like perception gets less reliable when the object of perception is further away, so too get epistemologies of possibility less reliable when they consider 'remote' possibilities. The latter argues that if one focuses on an epistemology of possibility, then we cannot get to the relevant level of detail in order to justify beliefs in certain possibility claims. Both arguments come down to similar issues (that of relevant-depth) and both concern the limits of our *imaginative capacities*. Ignoring the fact that there are other empiricist epistemologies of possibility, this is ultimately a motivation for modal modesty that rests on the limits of the cognitive capacities of *ordinary humans*. It is ordinary agents, and not idealised Laplacean demons, that fail to imagine the relevant details. In that sense, these motivations are subsumed in the ordinary cognition motivation.

[12] Strohminger and Yli-Vakkuri reject a kind of evolutionary explanation that Nichols favours.

11.5 Consequences of Modal Modesty

Let me conclude by noting some potential consequences of accepting a form of modal modesty.[13] First of all, accepting modal modesty raises the worry about where to draw the line between those instances where we can be said to have justified beliefs in what is possible and cases where we cannot. For example, Leon (2017, p. 253) argues, against Van Inwagen's theory, that if the drawing of the line is unprincipled and *ad hoc*, then this line will become 'unstable' (Geirsson, 2005 also explicitly develops this challenge to moderate modal scepticism, as does Hawke, 2011, who sets out to rebut it). The worry is that one's modal modesty might collapse into more radical forms of scepticism: if the line between cases of which we can know their modal status and those of which we cannot is unprincipled, then why are the motivations for modal modesty not simply motivations for radical modal scepticism? This, I think, is indeed a fair challenge to proponents of modal modesty and it is not obvious that all motivations (e.g., Van Inwagen 1998) are in the clear with regards to it. However, the modally modest epistemology of possibility developed in Chapter 8 seems to rebut this worry (this kind of modal modesty is mainly motivated along the lines of the Ordinary Cognition motivation and is along the similar lines as Hawke's (2011) rebuttal). It simply is not obvious whether our ordinary cognitive capacities are reliable when judging more exotic hypothetical situations, but from this we shouldn't conclude that they are generally unreliable. In general, our ordinary cognitive capacities are very reliable. Surely, where exactly this motivation draws the line between those cases we can know to be possible and those we cannot may be unclear, but it does not collapse into radical modal scepticism.

Secondly, Machery argues that accepting modal modesty means that a "large swath of traditional and contemporary philosophy [...] must be abandoned" (2017, p. 187). And it seems true that modal modesty has an effect on our philosophical theorising. For example, arguments based on premises involving the possibility of exact physical duplicates without consciousness should be viewed with suspicion, as modal modesty suggests that it is not clear that we can know the truth of such a premise. Though, as argued in Chapter 10 (Section 10.5), this does not mean that all of (traditional) philosophical theorising should be shelved.

Additionally, there are some interesting open questions that deserve close attention in future work in the epistemology of modality. First of all, it is an open question how the kind of modal modesty discussed in this chapter relates to Roca-Royes' (2007; 2017) theory, which we might dub *moderate modal agnosticism*. Roca-Royes points out that her "position with respect to the knowability of the remote [cases] is, although congenial to his, a bit weaker than Van Inwagen's in that, where he is

[13] Hawke (2017) argues that accepting modal modesty, which he calls 'moderate modal scepticism', helps us disarm some more vicious forms of scepticism, such as Humean scepticism about induction. This suggests, Hawke concludes, that modal modesty has significant "*theoretical utility*" as it serves "as an *antidote to paradox*" (p. 304, original emphases).

sceptical, I am at the moment agnostic" (Roca-Royes, 2007, p. 119; see also Roca-Royes, 2017, p. 242). That is, one does not judge that we cannot know the modal status of, e.g., the claim that there are philosophical zombies, one merely withholds judgement about whether we can know such claims. Moderate modal agnosticism, as I take Roca-Royes to understand it, allows for the fact that "[f]or all that has been developed [in epistemologies of possibility], some such (perhaps-)truths [concerning remote metaphysical possibilities] might still be knowable *somehow else*" (2017, p. 242, original emphasis). Future work should investigate the points of agreement and disagreement between Roca-Royes' view and different kinds of modal modesty.

Secondly, remember from Chapter 1 (Section 1.4.1) that there is one issue for empiricist epistemologies of modality that we set aside: the integration challenge (Sjölin Wirling, 2019a; Roca-Royes, forthcoming). This is the issue of providing a modal metaphysics that makes the proposed epistemologies of modality credibly do justice to the modal metaphysics (Roca-Royes, forthcoming, p. 2). Accepting modal modesty might provide some relief here. For example, Mallozzi (2018a) argues that we should focus on a *metaphysics-first* approach to modality, as "we cannot hope to explain how we know the truths of a given domain without some conception of what constitutes the truths of that domain" (pp. 1-2). Yet, if modal modesty is correct, then we might also not be in a position to (fully) determine what constitutes the modal truths. The kinds of constitutional facts that are often suggested (i.e., facts about essences) are exactly the kinds of facts that modal modesty suggests we might not be able to form reliable judgements about. Investigating exactly what the consequences are of accepting modal modesty with regards to the integration challenge, is something that deserves close attention in future work.

Detailed evaluation of the different varieties of modal modesty, the possible motivations for it, and these open questions is left for future work. The work in this dissertation suggests that there are promising epistemologies of possibility that are cognitively plausible and methodologically naturalistic. These epistemologies explain how we can have justified beliefs in possibility claims, most notably mundane, ordinary ones. Based on this aim for cognitive plausibility, I think we should accept modal modesty. How exactly this influences our epistemology of modality and our philosophical theorising is something that I can only hint at at this point. A tentative conclusion is that a promising and prominent naturalistic programme, spanning traditional and experimental approaches to philosophy, is plausibly committed to *both* the reliability of possibility judgements, some of philosophical import (e.g., typical Gettier-reasoning), *and* modally modest philosophy.

Appendices

Appendix A
Representational Imagination

In this appendix,[1] I will elaborate on two imagination-based epistemologies of possibility within the representational view of imagination (in addition to the QALC imagination discussed in Chapter 3): one focusing *solely* on the qualitative content of imagination and one allowing *unrestricted* linguistic content. Prospects for both of these are rather bleak, which, for many, is a motivation to turn to more sophisticated views of representational imagination, such as theories of QALC imagination discussed Chapter 3.

A.1 Imagination as Purely Pictorial

When thinking about the role of imagination in the epistemology of possibility, it might seem natural to suggest that only imaginings with *purely qualitative* contents should count as evidence for possibility.[2] Imagination – or at least those imaginings that feature in the epistemology of possibility – on such an account should represent a situation without any element of language-like, arbitrary labelling, or meaning

[1] The material of this appendix is based on Sections 3 and 4 from Berto & Schoonen (2018). Additionally, see Berto & Schoonen (2018, sec. 2) and Thomas (2018, especially the 'Dual Coding and Common Coding Theories of Memory' supplement) for discussions of the dual versus common coding and propositional versus analog debates in the philosophy of mind respectively. These debates, both emerging in the 70s and 80s of the previous century, had a huge impact on the field of philosophy of mind, forcing the emphasis of much of the ensuing research towards mental content in its qualitative or linguistic form. Interestingly, Thomas (1999, 2018) argues that it is because of these debates and their impact that research approaching mental faculties from perspectives *other than* the representational content only gained popularity much later. See Chapter 5 for a discussion of a theory of imagination that is compatible with views sceptical of mental representations such as Chemero (2009) and Hutto & Myin (2012).

[2] Hume (1777/1997) held something close to an epistemology of possibility based on purely qualitative content (see Lightner, 1997; Kail, 2003; van Woudenberg, 2006; and Dohrn, 2010). See Kung (2017) for an elaborate discussion of such a 'Humean' account, its limitations, and the move to contemporary imagination-based epistemologies of possibility.

assignment (i.e., no linguistic content), but rather, purely qualitatively: only via the phenomenological and quasi-spatial similarity of the imagery to the situation or world making φ true.

This is rather demanding. For note that even physical pictorial representations need not represent purely qualitatively. If one makes a drawing representing a river flowing east to west and a tree with a round-shaped crown of leaves north of the river, by having a blue line running from the left to the right of the sheet, and above it a shorter brown line oriented bottom–up, with a green circle on top of it. This drawing represents what it represents, partly by chromatic and geometric similarity between the coloured areas of the sheet and the shape and colour of the tree and of the river, and partly by the stipulation that x's being north of y be represented by the representation of x's being drawn above that of y on the sheet. Even more importantly, it represents via the stipulation that the green patch with a brown line below it represents *the tree*, and the blue line oriented from left to right on the sheet represents *the river*. Some have even argued that mental representation can *never* work in cognition only by similarity, or purely qualitatively: from Goodman's (1976) general argument opposing the symmetry of similarity to the asymmetry of the representation relation, to Fodor's charges of lack of compositionality and insufficient specificity for mental imagery (Fodor, 1975, 1981).

Yet, for the sake of the argument, assume that these criticisms don't work; that there does exist mental imagery representing purely qualitatively, with no labelling or arbitrary meaning-assignment; and that this makes for the kind of imagination involved in imagination-based epistemologies of possibility. The only scenarios imaginable in this way seem to be those that involve exclusively primary and secondary perceptual qualities (colours, shapes, extension, motion) of physical objects arranged in space-time. As many have pointed out, if this is *all* that we have to go on in our imaginations, then there are not many possibilities we can get justification for through imagination. For example, the fact that Susan and Andy could be friends, the fact that today could be Friday, the fact that Quinn and Blake could be second cousins; all of these situations involve non-qualitative properties and thus knowledge of their possibility can *not* be accounted for by the purely qualitative content of imagination. We can never imagine, in the relevant sense, situations involving abstract objects, or any non-perceptual feature of concrete objects. Kung (2016) points out that "[i]n fact, [we] can't specify *anything* about the thing's constitution *without assignments*" (p. 113, second emphasis added). That is, we cannot imagine that a table could be made out of wood based on purely qualitative content alone.[3] Additionally, for any non-actual possible (non-)identity we need linguistic content. E.g., it seems plausible that Obama could have a third daughter, call her 'Michelle Jr.', and that she would be distinct from Obama's firstborn, Malia Ann Obama. All of these imaginings cannot justify the relevant propositions based solely on their

[3] Note that the indefinite article, 'a', is crucial here in order to avoid issues with the Kripkean essentiality of constitution.

qualitative content (see also Kung, 2010, 2017).

So, purely qualitative imagination cannot justify our knowledge of possible non-qualitative properties, constitutional properties, and identities. This leaves an incredibly small set of possibility claims where such imagination would be able to justify our knowledge of these possibilities. As Kung puts it, "purely imagistic imaginings comprise a *very small* subset of imaginings" (2017, p. 136, emphasis added). The result would be a form of modal scepticism that goes far beyond that of moderate modal sceptics such as Van Inwagen (1998) and Hawke (2011); especially because many very mundane, ordinary possibility claims would not be able to be justified through imagination thusly understood (e.g., any possibility involving *me*, rather than a qualitative duplicate). Remember that these basic modal claims are such that, according to Hawke, "a theory of modal epistemology or modal metaphysics is likely to be viewed with suspicion if it suggests that we are *not justified in believing* [them]" (2011, p. 360, emphasis added).

A.2 Imagination with Unrestricted Linguistic Content

Given that imaginings without linguistic content seem to be too weak for a significant epistemology of possibility, let us turn to the other end of the spectrum: allowing *unrestricted* linguistic content.

Suppose that linguistic mental contents have at least the same representational power as the expressions of natural languages like English. Call this the Parity Assumption: whatever content is representable by a natural language sentence, is also representable by some linguistic mental content. If linguistic (mental) content is understood just as natural language sentences tokened in the head, the Parity Assumption is obvious. But even if one claims that the relevant content is more deeply encoded, say, in a (by hypothesis, unconscious) Fodorian language of thought (Fodor, 1975), one should grant that whatever content can be represented in natural language can also be represented in mentalese, given that the latter is (again, by hypothesis) supposed to ground the learnability and mastering of the former.

If the Parity Assumption is right, then it is very plausible that, given unrestricted linguistic content, we can imagine the impossible. To deny it, one would seem to be forced to make one of two moves: (1) claim that sentences of ordinary languages like English, describing alleged absolute impossibilities, actually are meaningless strings. Or, (2) claim that although those sentences are meaningful, and so by the Parity Assumption we can have corresponding, meaningful, linguistic mental representations, we cannot understand these.

But claim (2) is incredible in the face of the compositionality of learnable languages. Let p be any simple, intelligible sentence of English, such as 'This table is round'. Surely p cannot become unintelligible because we stick a negation in front of it. So $\neg p$ must be intelligible, too. And surely two such sentences cannot deliver an unintelligibility once we conjoin them, $p \wedge \neg p$. So the latter must be intelligible, too, and by the Parity Assumption we can have a corresponding linguistic mental representation which will be intelligible in its turn, and whose content is *that* $p \wedge \neg p$. But (unless one is a dialetheist: see Priest, 1998; Priest et al., 2018), contradictions are true in no possible world.

So we are left with claim (1). Someone who came close to making it is Wittgenstein (1922). 'Came close', because for Wittgenstein's *Tractatus* tautologies, logical truths, and their negations, logical falsities, are notoriously *sinnlos* (4.461). They "say nothing" (ibid.). Even for Wittgenstein they "are, however, not senseless [*unsinnig*]" but "part of the symbolism in the same way that '0' is part of the symbolism of arithmetic" (4.4611). There is a debate among Wittensteinians, on what the difference between *sinnlos* and *unsinnig* amounts to, but we don't need to get into this. One straightforward interpretation of the Wittgensteinian view, in the contemporary terminology of possible worlds, is that the informative job of a sentence is to split into two the totality of worlds: those in which the sentence is true and those in which it is false. The former group gives the proposition expressed by the sentence in standard possible worlds semantics. But then tautologies and their negations, being true everywhere and nowhere in the modal space respectively, don't split, and turn out to be uninformative: "I know, e.g., nothing about the weather, when I know that it rains or it does not rain" (4.461).

Even if one buys the view that logical truths and falsities are uninformative,[4] one need not accept that this makes them *contentless*. Quine (1948) makes the point of the meaningfulness of contradictions in 'On What There Is', as a response to fictional philosopher Wyman, sometimes taken as representing Meinong's view that some things do not exist (see Berto, 2013). Wyman believes that things like Pegasus ought to be admitted in our ontological catalogue, as *possibilia*, for otherwise it would make no sense to even say that Pegasus is not. By parity of reasoning, objects Quine, we ought to admit the round square cupola on Berkeley College; otherwise, it would make no sense to even say that *it* is not. But accepting this brings inconsistency. Wyman reacts by declaring that inconsistent conditions are meaningless (i.e., contentless). Quine's reply is spotless:[5]

[4] Which, on its own, seems like an implausible view. Consider what can be learned by a rational, but finite and fallible agent – one of us. We can learn that a complex formula, whose truth value we were ignorant of until we computed its long truth table, is a tautology. For all we knew before carrying out the computation, the formula's being false was a way things could be. In this sense, *pace* Wittgenstein (6.1251), there are surprises in logic. A book defending this view is Jago (2014).

[5] Priest, who accepts true contradictions, agrees:

> If contradictions had no content, there would be nothing to disagree with when someone uttered one, which there (usually) is. Contradictions do, after all, have meaning. If they did not, we could not even understand someone who asserted a

> Certainly the doctrine [of the meaninglessness of contradictions] has no intrinsic appeal; and it has led its devotees to such quixotic extremes as that of challenging the method of proof by *reductio ad absurdum* – a challenge in which I sense a *reductio ad absurdum* of the doctrine itself.
>
> Moreover, the doctrine of meaninglessness of contradictions has the severe methodological drawback that it makes it impossible, in principle, ever to devise an effective test of what is meaningful and what is not. It would be forever impossible for us to devise systematic ways of deciding whether a string of signs made sense – even to us individually, let alone other people – or not. For it follows from a discovery in mathematical logic, due to Church (1936), that there can be no generally applicable test of contradictoriness. (1948, pp. 34-35)

One may still object as follows.[6] In the view under attack, imagining φ is bearing a certain relation, call it I, to a linguistic mental representation S, which means *that* φ. In a familiar metaphor, it is to have a representation S which means that φ in one's 'imagining box'. Granted, there are impossible linguistic representations (by the Parity Assumption). But can we bear I to them? Surely a corresponding 'belief box' model should not be committed to the view that a cognitive agent, x, can believe (have a 'B-relation' to) an impossibility, just as it shouldn't be committed, say, to the believability by x that x itself does not exist. Even when 'I do not exist' is (suppose) a meaningful mentalese sentence, that doesn't mean it can be in x's belief box when 'I' picks out x. As an analogy, take a bulletin board on which announcements can be pinned. It may be a well-enforced rule that no political flyers can be attached to the board even though the content of the flyers is perfectly meaningful and intelligible. Something could prevent 'I do not exist' from being in one's belief box; and similarly for the imagination box.

Even if one accepts the boxology terminology, one should resist the analogy between imagination (with unrestricted linguistic content) and belief. The boxology terminology is supposed to suggest that, e.g., belief and imagination can operate on *the same kind* of objects – i.e., whatever one takes mental content to be. However, all agree that belief and imagination are still *functionally* different in significant ways (e.g., Currie & Ravenscroft, 2002; Nichols & Stich, 2003; Langland-Hassan, 2016). In particular, as for example argued by Langland-Hassan (2016), imagination is subject to voluntary control in ways believing is not. Conscious acts of imagination often have an *arbitrary* starting point. This may be made up by the agent ('Now let's

> contradiction, and so evaluate what they say as false (or maybe true). We might not understand what could have brought a person to assert such a thing, but that is a different matter and the same is equally true of someone who, in broad daylight, asserts the clearly meaningful 'It is night'. (1998, p. 417)

[6] Thanks to an anonymous referee of (Berto & Schoonen, 2018) for bringing up this point. The objection is phrased as in their original comment.

imagine what would happen if. . . '), or it may be given as an external instruction (think of going through a novel, taking the sentences you read as your explicit input as you revise your imagined scenario).[7]

Nichols (2006a) explicitly points out that our belief-box rejects contradictory representations: "[y]ou don't need to be much of an evolutionary psychologist to agree that it would be adaptive for animals to stop believing p when they come to believe $\neg p$" (p. 251). But, given the functional differences between imagination and belief, there is no reason to think that imagination with unrestricted linguistic content is similarly restricted. A plausible explanation for why this is so is that imagination – again, understood as the having unrestricted linguistic content – is neutral in ways believing is not: believing requires commitment, which is absent when one just imagines (see Balcerak Jackson, 2016 on this point with respect to *supposing*). Similarly, it would be pragmatically inconsistent to assert 'I do not exist', but it is not pragmatically inconsistent to consider the claim as an imagination ('Imagine my parents had never met, so I was never born; then this dissertation would not have been written. . . '). The attitude is one of allowing a certain content to show up for consideration, not taking a stance on its being realised.

In their influential book on mental simulation and imagination, Nichols & Stich (2003) make the point explicitly in terms of mental boxes. For Nichols and Stich, imagination works via what they call a 'possible worlds box', where we voluntarily put "an initial premiss or set of premisses, which are the basic assumptions about what is to be pretended" (p. 24). This box, for Nichols and Stich, is connected to our 'belief box' because we integrate the explicit pretense's content with a selection of our beliefs. However, they make clear that the two do not coincide and ought not to be confused, for we can bear the I-relation to lots of things we cannot bear the B-relation to. And in spite of their speaking of 'possible worlds', the explicit premise that makes for the starting point of our acts of mental simulation can well be impossible:[8]

> We are using the term 'possible world' more broadly than it is often used in philosophy [. . .], because we want to be able to include descriptions of worlds that many would consider *impossible*. For instance, we want to allow that the Possible World Box can contain a representation with the content *There is a greatest prime number*.
>
> (2003, p. 28, fn. 5, original emphases)

Thus, if imagination is understood merely as the having of unrestricted linguistic contents, we can imagine the impossible. In fact, as Hill has remarked, in this

[7] See Chapter 4 for more on this.

[8] Nichols (2006a) thinks that we cannot imagine *explicit* contradiction, however, he does think that we can imagine impossibilities. Moreover, his theory of imagination is more sophisticated than imagination with unrestricted linguistic content (for something that comes closer to his preferred view, see Chapter 4).

sense – which he dubs "simple, undisciplined conceiving" – "virtually anything is conceivable", and "conceivability is therefore incapable of providing a reliable test for possibility" (Hill, 2016, p. 326).

A.3 An Impasse

We seem to have reached an impasse: imaginations based on purely qualitative content are too weak to play a significant role in an epistemology of possibility, while allowing in unrestricted linguistic content opens the gates to imagining all sorts of impossibilities.

QALC theories of imagination aim to overcome this impasse, as we saw, by allowing in only authenticated linguistic content. I discuss these in Chapter 3.

Appendix B
Adding Topicality to Models of Pretense Imagination

In this appendix, I will enrich the models of pretense imagination proposed in Chapter 4 with a *topicality* component. This will help us overcome the idealisations imposed by the former framework and shortcomings of some previous logics of imagination. Additionally, a worry for the logic of imagination of Berto (2018a,b) is discussed. I show that the model presented here is sufficiently rich to overcome this issue and, thus, provides us with a further step in the right direction toward developing an adequate formalisation of imagination in pretense.

B.1 Topicality in Pretense

In Chapter 4 (Section 4.1), I discussed how pretense imagination relates to beliefs and, in particular, how the belief-like reasoning and the background beliefs of an agent restrict the development of pretense imagination. There, I already hinted (in footnote 6) that there is an additional component that restricts the development of pretense imagination: *topicality.* The way to see this is that even though agents engaging in pretense imagination take on board some background beliefs (e.g., about tea-party), it seems obvious that the agent does not take all their background beliefs on board. Why is it that some other background beliefs, such as Paris being the capital of France, water being a transparent liquid, etc., are not taken on board? I argue that one of the reasons why the subject does not imagine Paris being the capital of France in the tea-party situation is simply that the capital of France is *off-topic* and *irrelevant* to the pretend tea-party. This suggests a natural way to separate the background beliefs that can be taken on board in the pretense from the ones that are not: we select the *relevant* background beliefs to import into pretense based on what they are about, in other words, based on their *topics* (see Berto, 2018a,b for aboutness in imagination).

In this section, I will first discuss the notion of topics and aboutness more general, after which I will raise the issue of imaginative episodes having an *overall topic*. Then, I will briefly highlight some of the idealisations of the formal models of Chapter 4 (Section 4.3). This motivates adding topic-models, which I will do in Section B.2.

B.1.1 Aboutness: topicality

In a series of work, Berto (2018a,b), Berto & Hawke (2018), and Hawke (2018) have developed a general theory of *topic-sensitive* propositional content, which has also been used in epistemic contexts to address problems of logical omniscience (Berto, 2019; Berto & Özgün, 2020; Hawke et al., 2020). I briefly recap the main components of their proposal, but refer to the aforementioned sources for a more detailed presentation.

Within pretense imagination, we focus only on *propositional* imagination: imagining *that* such and such is the case (see Chapter 2, Section 2.1.1). Imagination, as a mental attitude towards propositions, thus ranges over propositional contents, which are generally taken to be sets of possible worlds. However, treating propositional content this way leads to too crude an identification of propositions that causes serious idealisation problems – such as the problems of logical omniscience – for formal representations of mental attitudes. Here is an example. Since 'Extremally disconnectedness is not a hereditary property of topological spaces' and 'Jane is a logician or she is not a logician' are true at exactly the same (namely, all) possible worlds, they are treated to represent the same proposition. However, they obviously do not say the same thing as they differ in *topic* (indicated by **boldface**): the latter is about **Jane, Jane's profession** etc., whereas only the former is about, e.g., **extremally disconnectedness, hereditary properties, topology** but not about **Jane**. One can grasp facts about Jane without having even heard of what a topology is. So, arguably, we can imagine, believe, know the latter without imagining, believing, knowing the former and *vice versa*. While this is difficult to represent (if possible at all) by using *only* the standard possible worlds semantics and Hintikka's (1962) way of modelling (propositional) mental attitudes as quantification over possible worlds, supplementing the standard possible worlds semantics with an account of *aboutness* – "the relation that meaningful items bear to whatever it is that they are on or of or that they address or concern" (Yablo, 2014, p. 1) – solves the problem to a great extent (see, e.g., Berto, 2018a,b; Berto & Hawke, 2018; Hawke et al., 2020).[1] The content of an interpreted sentence then becomes a pair of its (1) intension and (2) topic. Thus, in particular, imagining a proposition requires also knowing what it is about, i.e., *having grasped* its topic.

It is consensus in theories of partial content that truth functional logical connectives do not add anything to the topic of a sentence, that is, they are *topic-transparent* (Fine, 1986, 2016; Hawke, 2018). Whatever is on topic with 'Jane is a

[1] The problem of the overall topic remains; as we will see below.

logician' is also on topic with 'Jane is *not* a logician' and *vice versa.* They are about exactly the same things, e.g., **Jane** and **Jane's profession**. Similarly, the topic of 'Jane is a logician and Kate is a philosopher' is the same as that of 'Jane is a logician or Kate is a philosopher'. It is a *fusion* of the topics of 'Jane is a logician' and 'Kate is a philosopher'. Additionally, the topic of 'Kate is a philosopher' is *part of* the topic of 'Jane is a logician and/or Kate is a philosopher'. That is, topics can be fused together and include other topics as their proper parts. They stand in a *mereological relation.* All of this will be reflected in the formal models developed in Section B.2.

Overall Topics

Using topicality in formal models of imagination is done by, e.g., Berto (2018a,b), who presents a formalisation of propositional imagination that incorporates a topicality component that represents the topic-sensitivity of (propositional) mental states. While his logics of imagination successfully employ (conditional) modal operators that can discern logically and necessarily equivalent propositions, they fall short of representing the *overall topic* of an imaginative episode, an important factor affecting the development of pretense imagination. To see what I mean by 'overall topic' and how this affects the imagination, consider the following two situations:

> **Context A:**
> You are flying to Australia the day after tomorrow to take a well-deserved holiday. That evening, when watching the news, you find out that there is a tornado in Indonesia and that nothing else is known at this point. You wonder whether this influences your flight.
>
> **Context B:**
> You have a friend living in Singapore, who lives right by the coast. That evening, when watching the news, you find out that there is a tornado in Indonesia and that nothing else is known at this point. You wonder whether this might affect your friend.

In order to help you evaluate the effects of the tornado in each case, you engage in an imaginative exercise. In particular, in both cases, you use the following explicit input

(9) There is a tornado in Indonesia

and start the imaginative process to determine the effects of the tornado. As **Context A** involves holiday planning and **Context B** is concerned with your friend living close to a tornado zone in Indonesia, the imaginings resulting from (9) could be different in **Context A** and **Context B**. For example, imagining 'Booking a flight through the US rather than Indonesia is safer' seems to be *off-topic* in **Context B**, whereas it is *on-topic* in **Context A**.

The above example is no exception, imagination is often influenced by its overall topic. So a formal model of imagination should be able to account for the fact that different contexts – based on the exact same explicit input and background beliefs – might give rise to different imaginative episodes solely due to their distinct *overall topics*.[2] Berto's (2018a; 2018b) logics of imagination, however, are unable to do so, as these logics only focus on the relationship between the topics of particular input/output propositions and overlook the idea that there might be overall topics to exercises of imagination. I suggest a way forward, by not only focusing on the topic of the particular propositions involved, but also adding, what I will call, an *overall topic* to our model of pretense imagination.[3]

B.1.2 Idealised Imaginers

The fact that pretense imagination seems to be restricted by the topic of the imaginative episode is a philosophical motivation to add topic models. Additionally, the framework as it is described in Chapter 4 (Section 4.3; especially Definitions 4 and 5, on pages 70 and 72 respectively) results in highly *idealised* imaginers.

Let me explain by saying something about the logical properties of the modal operators B and I. As standard for the belief modality interpreted as *truth in the most plausible worlds*, our agent believes all logical truths and their beliefs are closed under believed implications:[4]

Omniscience rule for B: if $\vDash \varphi$, then $\vDash B\varphi$

Closure under believed implications: $\vDash B(\varphi \Rightarrow \psi) \Rightarrow (B\varphi \Rightarrow B\psi)$

As a consequence of the above principles, it is also the case that the agent believes all logical consequences of what they believe and their beliefs are closed under logical equivalences:

Closure under valid implications for B: if $\vDash \varphi \Rightarrow \psi$, then $\vDash B\varphi \Rightarrow B\psi$

Closure under valid equivalences for B: if $\vDash \varphi \Leftrightarrow \psi$, then $\vDash B\varphi \Leftrightarrow B\psi$

[2] The notion of an overall topic is inspired by an objection raised against Berto's work by Timothy Williamson when the former presented some of his work at the 'Philosophy of Imagination' conference at the Ruhr Universität Bochum in March 2018.

Independently of the work on which this Appendix is based, Canavotto et al. (2020, sec. 2) suggest that imagination is *goal-driven*, which is related to the notion of overall topic used here.

[3] The particular form of the models is not essential to this enrichment. The same solution could also be implemented in Berto's (2018a; 2018b) models of imagination.

[4] Remember that here we use '⇒' for the material implication.

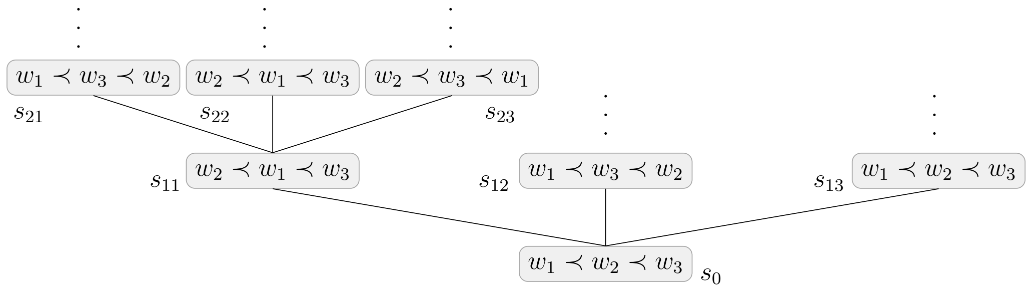

$w_1 \prec w_3 \prec w_2$
$w_2 \prec w_1 \prec w_3$
$w_2 \prec w_3 \prec w_1$
s_{21}
s_{22}
s_{23}
s_{11}
$w_2 \prec w_1 \prec w_3$
s_{12}
$w_1 \prec w_3 \prec w_2$
s_{13}
$w_1 \prec w_2 \prec w_3$
$w_1 \prec w_2 \prec w_3$
s_0

Figure B.1: *Counterexample 1. The plausibility ordering of each stage is given in the corresponding box.*

The agent in question is therefore highly idealised, in the sense that they are logically omniscient with respect to their beliefs.

The imagination operator I on the other hand is weaker. The agent does not necessarily imagine all logical truths, their imagination is not closed under imagined implications and they do not necessarily imagine all logical consequences of what they imagine, i.e., the following *fail*:

Omniscience rule for I: if $\vDash \varphi$, then $\vDash I\varphi$

Closure under imagined implications: $I(\varphi \Rightarrow \psi) \Rightarrow (I\varphi \Rightarrow I\psi)$

Closure under valid implications for I: if $\vDash \varphi \Rightarrow \psi$, then $\vDash I\varphi \Rightarrow I\psi$

Counterexample 1: Consider the model $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, V\rangle$ in Figure B.1, where $W = \{w_1, w_2, w_3\}$ such that $V(q) = \{w_1\}$ and $V(p) = \{w_2\}$. The rest of the model is as depicted in Figure B.1 and, for the sake of this argument, it is sufficient to focus on the branches that include stage s_{11}. For omniscience rule for I: $p \vee \neg p$ is a logical validity, but $\langle w_1, (s_0, s_{11})\rangle \nVdash I(p \vee \neg p)$ since $\preceq_{s_{11}} \neq \preceq_{s_0}^{(p\vee\neg p)} = \preceq_{s_0}$. Moreover, for closure under valid implications, we have $p \Rightarrow (p \vee \neg p)$ logically valid and $\langle w_1, (s_0, s_{11})\rangle \Vdash Ip$ since $\preceq_{s_{11}} = \preceq_{s_0}^{p}$ and $\langle w_1, (s_0, s_{11})\rangle \Vdash Bp$. However, $\langle w_1, (s_0, s_{11})\rangle \nVdash I(p \vee \neg p)$ as shown above. As a counterexample for closure under imagined implications, consider the world-history pair $\langle w_1, (s_0, s_{11}, s_{21})\rangle$: we have $\langle w_1, (s_0, s_{11}, s_{21})\rangle \Vdash I(p \Rightarrow q)$ since $\preceq_{s_{21}} = \preceq_{s_{11}}^{(p\Rightarrow q)}$ as $|p \Rightarrow q| = \{w_1, w_3\}$, and $\langle w_1, (s_0, s_{11}, s_{21})\rangle \Vdash B(p \Rightarrow q)$. Moreover, $\langle w_1, (s_0, s_{11}, s_{21})\rangle \Vdash Ip$ since $\preceq_{s_{11}} = \preceq_{s_0}^{p}$ and $\langle w_1, (s_0, s_{11})\rangle \Vdash Bp$. However, $\langle w_1, (s_0, s_{11}, s_{21})\rangle \nVdash Iq$ since the sequence (s_0, s_{11}, s_{21}) cannot be obtained via an upgrade by q. That it, $\preceq_{s_{11}} \neq \preceq_{s_0}^{q}$ and $\preceq_{s_{21}} \neq \preceq_{s_{11}}^{q}$.

Note, however, that if φ and ψ are logically or necessarily equivalent, imagining one automatically leads to imagining the other. In other words, the following principle *does* hold in full models:

Closure under valid equivalences for I: if $\vDash \varphi \Leftrightarrow \psi$, then $\vDash I\varphi \Leftrightarrow I\psi$

This is because the belief revision method in place – the lexicographic upgrade – cannot distinguish logically or necessarily equivalent propositions: $\preceq_s^{\varphi} = \preceq_s^{\psi}$ if $\models \varphi \Leftrightarrow \psi$. Therefore, although weaker than belief, the operator I still renders the agents unrealistically idealised with respect to their imagination. For example, according to the proposed semantics, if the agent imagines at a stage that Jane is a logician or she is not, they also imagine that $2 + 2 = 4$. Intuitively, we can imagine or believe the former without imagining or believing the latter and *vice versa*. In addition, while the former might be on-topic with an imaginative episode about Jane, the latter is not necessarily so. In a similar vein, consider again the tea-party example from Chapter 4 (Section 4.1). The agent does imagine that one of the cups is full, however, they do not imagine that one of the cups is full *and 2 + 2 = 4*, even though these two sentences are logically equivalent. Moreover, they do not import their irrelevant beliefs about Paris being the capital of France to the imaginative episode as these are completely off-topic to the imaginative episode in question. The model proposed in Chapter 4 (Section 4.3) is unable to account for such cases.

B.2 What's it all About: Adding Topicality

This section aims at refining the formal models of Chapter 4 (Section 4.3) in a way that the modal operators B and I become more sensitive to distinctions between logically equivalent contents. To do so we endow branching-time belief revision models with (an enriched version of) topic models introduced in Berto (2018a). This way we can evade the problems concerning the aforementioned idealisations.

Definition 8. **Topic Model for $\mathcal{L}$**
A *topic model* $\mathcal{T}$ is a tuple $\langle T, \oplus, t\rangle$, where

1. T is a finite, non-empty set of *possible topics*;

2. $\oplus : T \times T \to T$ is a binary idempotent, commutative, associative operation: *topic fusion*. We assume unrestricted fusion, that is, $\oplus$ is always defined on T: $\forall \mathbf{a}, \mathbf{b} \in T\ \exists \mathbf{c} \in T(\mathbf{c} = \mathbf{a} \oplus \mathbf{b})$;[5]

3. $t : \mathsf{Prop} \to T$ is a *topic function* assigning a topic to each element in Prop. t extends to the whole $\mathcal{L}$ by taking the topic of a sentence φ as the fusion of the topics of the atomic propositions occuring in it. I.e.,

$$t(\varphi) = \oplus \mathsf{AT}(\varphi) = t(p_1) \oplus \cdots \oplus t(p_n),$$

where $\mathsf{AT}(\varphi) = \{p_1, \ldots, p_n\}$ is the set of propositional variables occurring in φ.

[5] We take the set of topics to be finite. If one thinks that there are infinite topics, then one can

In the metalanguage we use variables $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ ($\boldsymbol{a_1}, \boldsymbol{a_2}, \dots$) ranging over possible topics. *Topic parthood*, denoted by $\sqsubseteq$, is defined in a standard way as

$$\forall \mathbf{a}, \mathbf{b}(\mathbf{a} \sqsubseteq \mathbf{b} \text{ iff } \mathbf{a} \oplus \mathbf{b} = \mathbf{b}).$$

Thus, $(T, \oplus)$ is a *join semilattice* and $(T, \sqsubseteq)$ a *poset*. The *strict topic parthood*, denoted by $\sqsubset$, is defined as usual as $\mathbf{a} \sqsubset \mathbf{b}$ iff $\mathbf{a} \sqsubseteq \mathbf{b}$ and $\mathbf{b} \not\sqsubseteq \mathbf{a}$.

The topic of a complex sentence φ is defined from its primitive components in $\mathsf{AT}(\varphi)$, where all the logical connectives, as argued in Section B.1.1, are topic-transparent. We therefore have that for all $\varphi, \psi \in \mathcal{L}$,

- $t(\neg\varphi) = t(\varphi)$
- $t(\varphi \wedge \psi) = t(\varphi \vee \psi) = t(\varphi \Rightarrow \psi) = t(\varphi \Leftrightarrow \psi) = t(\varphi) \oplus t(\psi)$.

Topic models provide an abstract and objective (i.e., agent independent) representation of the mereological structure of topics assigned to Boolean sentences and, in turn, help us make distinctions between logically equivalent contents (Berto, 2018a,b). However, as argued in Section B.1.1, Berto's theory is too coarse-grained in that it cannot account for the possibility that exactly the same explicit input can lead to different imaginative episodes due to their distinct overall topics (recall the example about the tornado in Indonesia). The reason why Berto's account is unable to deal with this issue is, I suggest, that his topic models include neither a representation of the overall topic of the imaginative episode nor the totality of topics the agent has mastered already (though the latter has been employed in recent work by Berto & Özgün, 2020; Hawke et al., 2020). Let's add these two components to the topic models in order to overcome the aforementioned issues. Another important operator, which we use towards the same purpose, is the so-called *topic intersection* $\sqcap : T \times T \to T$ such that $\mathbf{a} \sqcap \mathbf{b} = \oplus\{\mathbf{c} \in T : \mathbf{c} \sqsubseteq \mathbf{a} \text{ and } \mathbf{c} \sqsubseteq \mathbf{b}\}$. In words, $\mathbf{a} \sqcap \mathbf{b}$ is the fusion of all topics that are a common part of both $\mathbf{a}$ and $\mathbf{b}$.

We can now define a *topic-sensitive* version of branching-time belief revision models

Definition 9. **Topic-sensitive model**
A *topic-sensitive model* is a tuple $\langle S, \rightarrowtail, W, \preceq, T, \oplus, t, \mathfrak{b}, \mathfrak{i}, V \rangle$

1. $\langle S, \rightarrowtail, W, \preceq, V \rangle$ is a model;
2. $\langle T, \oplus, t \rangle$ is a topic model for $\mathcal{L}$;
3. $\mathfrak{b}$ and $\mathfrak{i}$ are designated elements of T such that $\mathfrak{b}$ represents 'the totality

also close T under arbitrary fusions. Either one of these things is needed in order to ensure that the topic fusion operator is well-defined.

of topics the agent has grasped' and $\mathfrak{i}$ represents 'the overall topic of the imaginative episode'.

A topic-sensitive model is equipped with a topic of the imagination exercise, $\mathfrak{i}$, and the totality of the topics the agent has grasped already, that is, $\mathfrak{b}$. These two components together impose a topicality filter on what the agent believes and imagines; thus resolving the issues noted at in Sections B.1.1 and B.1.2.

Component $\mathfrak{b}$ will make sure that the agent cannot believe those propositions whose topic they have not mastered yet. Intuitively, one does not believe that extremally disconnectedness is a hereditary property if they have never heard of, e.g., the topological properties 'extremally disconnectedness' or 'being hereditary'. Believing a proposition seems to require having a grasp of its topic. The designated element $\mathfrak{b}$ – i.e., the totality of topics the agent has grasped – allows us to account for this. Secondly, as mentioned above, one does not imagine everything they believe. Some of our beliefs might be off-topic with the given imaginative episode and a purposeful imaginative exercise seems to require keeping the imaginings within the subject matter of this imaginative episode. The component $\mathfrak{i}$ – i.e., the overall topic of the imaginative episode – helps to formally capture this. So, $\mathfrak{b}$ and $\mathfrak{i}$ together make it that pretense imagination is topic-restricted in two ways: the topic of what the agent imagines is a common part of both the totality of the topics the agent has grasped already *and* the overall topic of the imaginative episode. This idea is formalised by using the topic intersection operator $\sqcap$. These features will be better understood when looking at the new, topic-sensitive semantics for $\mathcal{L}_{\mathsf{BI}}$. While the semantics of the Booleans remain as they were before, the semantics of $B\varphi$ and $I\varphi$ are made stronger in the appropriate way with topicality constraints.

Definition 10. **$\Vdash$-Semantics for $\mathcal{L}_{\mathsf{BI}}$**
Given a topic-sensitive model $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, T, \oplus, t, \mathfrak{b}, \mathfrak{i}, V\rangle$ and world-history pair $\langle w, h\rangle$ such that $h = (s_0, s_1, \dots, s_n)$, the semantics for $\mathcal{L}_{\mathsf{BI}}$ is as given in Definition 4 (on page 70) for the components in $\mathcal{L}$, plus:

$$\begin{array}{lll} \mathcal{M}, \langle w, h\rangle \Vdash B\varphi & \text{iff} & Min_{\preceq_{s_n}}(W) \subseteq |\varphi|^h_{\mathcal{M}} \text{ and } t(\varphi) \sqsubseteq \mathfrak{b} \\ \mathcal{M}, \langle w, h\rangle \Vdash I\varphi & \text{iff} & \exists k < n(\preceq_{s_{k+1}} = \preceq^{\varphi}_{s_k} \text{ and } \langle w, h[k+1]\rangle \Vdash B\varphi) \\ & & \text{and } t(\varphi) \sqsubseteq \mathfrak{b} \sqcap \mathfrak{i} \end{array}$$

According to the topic-sensitive semantics, the agent believes φ at stage s iff (1) φ is true at all the most plausible worlds at s and (2) the agent has already grasped the topic of φ, i.e., the topic of φ is included in $\mathfrak{b}$ (see also Berto & Özgün, 2020). Therefore, the agent cannot believe φ within a pretense if they have not grasped its topic yet. Imagination, on the other hand, is restricted, additionally, by the overall topic of the imaginative exercise. The agent imagines φ if they have revised their belief state with φ at some earlier stage in the history and *the topic of φ is included in the intersection of the overall topic of the imaginative episode and the topic of* *the agent's belief state.*In topic-sensitive models with a singleton T, the semantics given in Definitions 4 and 10 coincide.

For the same reasons given in Chapter 4 (Section 4.3.1), the intended models are the 'full' version of topic-sensitive models. This time we also need to incorporate some topicality constraints.

DEFINITION 11. **Topic-sensitive full model**
A *topic-sensitive full model* $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, T, \oplus, t, \mathfrak{b}, \mathfrak{i}, V \rangle$ is a topic-sensitive model such that

1. for all $w \in W$, $h = (s_1, \ldots, s_n)$, and $\varphi \in \mathcal{L}$, if $\langle w, h \rangle \Vdash B\varphi$ and $t(\varphi) \sqsubseteq \mathfrak{b} \sqcap \mathfrak{i}$, then there is an $s' \in S$ such that $s_n \rightarrowtail s'$, $\preceq_{s'} = \preceq^{\varphi}_{s_n}$,
2. for all $s, s' \in S$, if $s \rightarrowtail s'$ then $\preceq_{s'} = \preceq^{\varphi}_{s}$ for some $\varphi \in \mathcal{L}$ such that $t(\varphi) \sqsubseteq \mathfrak{b} \sqcap \mathfrak{i}$.

The first condition states that whatever the agent believes is in principle available to be taken on board in the imaginative episode, as long as the belief is on topic with the overall topic of the imaginative episode. The second condition states that an agent revises their beliefs only according to the lexicographic upgrade policy and with only those propositions whose topics they have mastered and that fall under the overall topic of the imaginative episode.

The definitions of internally developed imaginative stages and intervened imaginative stages can be made topic-sensitive in a similar manner. I postpone a thorough study of these operators for another occasion. Let us now see to what extent the topic-sensitivity solves the aforementioned problems concerning idealisation and overall topic of an imaginative episode.

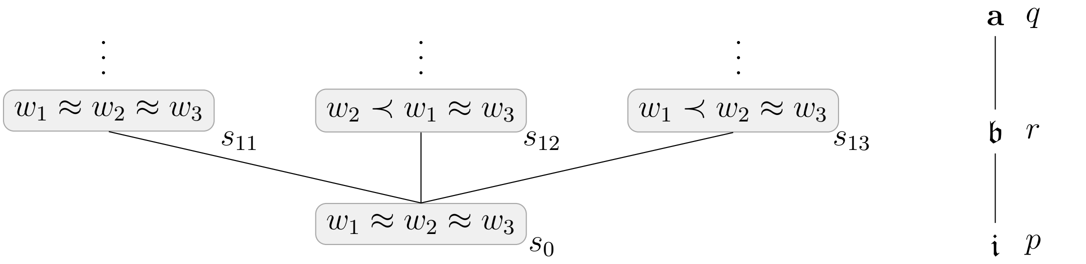
B.2.1 Idealisations, Tea-Parties, and Tornadoes

Topic-sensitivity allows us to model agents who do not believe all logical truths and whose beliefs are not closed under logical implications. That is, topic-sensitive full models *invalidate* the following principles:

Omniscience rule for B: if $\vDash \varphi$, then $\vDash B\varphi$

Closure under valid implications for B: if $\vDash \varphi \Rightarrow \psi$, then $\vDash B\varphi \Rightarrow B\psi$

Moreover, our agents can imagine/believe φ without imagining/believing ψ even when they are logically or necessarily equivalent.


$w_1 \approx w_2 \approx w_3$
s_{11}
$w_2 \prec w_1 \approx w_3$
s_{12}
$w_1 \prec w_2 \approx w_3$
s_{13}
$w_1 \approx w_2 \approx w_3$
s_0
$\mathbf{a}$ q
$\mathfrak{b}$ r
$\mathfrak{i}$ p

(a) $\langle S, \rightarrowtail, W, \preceq, V \rangle$ (b) $\langle T, \oplus, t, \mathfrak{b}, \mathfrak{i} \rangle$

Figure B.2: *Counterexample 2. The plausibility ordering of each stage is given in the corresponding box in Fig. B.2(a). Topic assignment is given by labelling the nodes in Fig. B.2(b) with atomic formulae.*

That is, the principles

Closure under valid equivalences for B: if $\vDash \varphi \Leftrightarrow \psi$, then $\vDash B\varphi \Leftrightarrow B\psi$

Closure under valid equivalences for I: if $\vDash \varphi \Leftrightarrow \psi$, then $\vDash I\varphi \Leftrightarrow I\psi$

no longer hold in topic-sensitive full models.

Counterexample 2: Consider the topic sensitive full-model $\mathcal{M} = \langle S, \rightarrowtail, W, \preceq, T, \oplus, t, \mathfrak{b}, \mathfrak{i}, V \rangle$ in Figure B.2, where $\langle S, \rightarrowtail \rangle$ and $\preceq$ are as given in Figure B.2(a), $W = \{w_1, w_2, w_3\}$, $T = \{\mathbf{a}, \mathfrak{b}, \mathfrak{i}\}$ with the topic lattice as depicted in Figure B.2(b). Finally, we consider three propositions p, q, r such that $V(p) = \{w_1\}$, $V(q) = \{w_1, w_2\}$, $V(r) = W$, and $t(p) = \mathfrak{i}$, $t(r) = \mathfrak{b}$, and $t(q) = \mathbf{a}$.

To refute closure under valid equivalences for I, let the actual history be $h = (s_0, s_{13})$. We then have $\langle w_1, h \rangle \Vdash Ip$, since $\preceq_{s_{13}} = \preceq^p_{s_0}$ and $t(p) = \mathfrak{i} \sqsubseteq \mathfrak{b} \sqcap \mathfrak{i} = \mathfrak{i}$. However, note that $\langle w_1, h \rangle \nVdash p \wedge (r \vee \neg r)$, since $t(p \wedge (r \vee \neg r)) = t(p) \oplus t(r) = \mathfrak{b}$ and $\mathfrak{b} \not\sqsubseteq \mathfrak{b} \sqcap \mathfrak{i} = \mathfrak{i}$. So, even though p and $p \wedge (r \vee \neg r)$ are *logically* equivalent, the agent can imagine the former without imagining the latter as r is *off-topic* with respect to the overall topic of the imaginative episode. This is exactly what we would expect. As a counterexample for the omniscience rule for B, take $\varphi := q \vee \neg q$, and for closure under valid implications and equivalences for B, consider $\varphi := p$ and $\psi := p \wedge (q \vee \neg q)$.

Tea-Parties and the Capital of France

Let us now stipulate that r:= Paris is the capital of France. In the model $\mathcal{M}$ given in Figure B.2 and every world-history pair (w, h) of $\mathcal{M}$, we have that $\langle w, h \rangle \Vdash Br$ (since $t(r) = \mathfrak{b}$ and $|r|^h_{\mathcal{M}} = W$). However, $\langle w, h \rangle \nVdash Ir$ since $t(r) = \mathfrak{b} \not\sqsubseteq \mathfrak{b} \sqcap \mathfrak{i} = \mathfrak{i}$, i.e., r is not on-topic with the modelled imaginative episode.

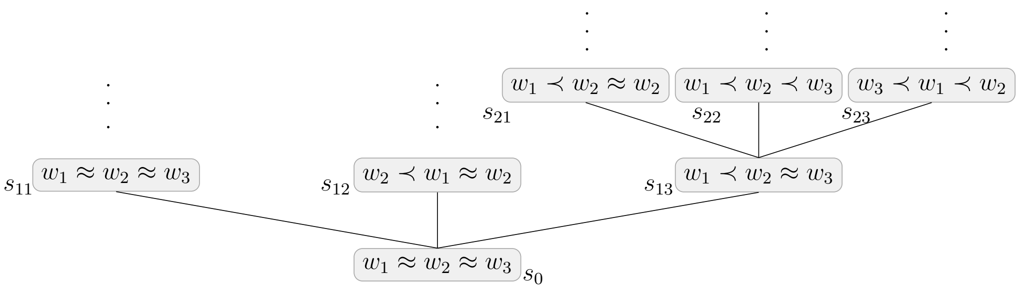
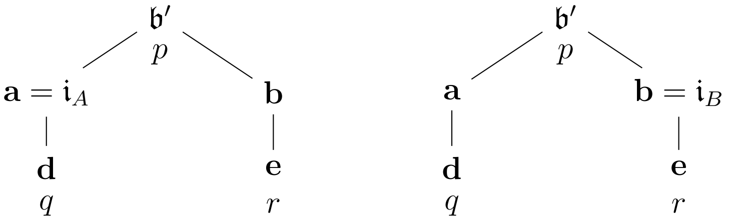

$w_1 \prec w_2 \approx w_2$
$w_1 \prec w_2 \prec w_3$
$w_3 \prec w_1 \prec w_2$
s_{21}
s_{22}
s_{23}
s_{11}
$w_1 \approx w_2 \approx w_3$
s_{12}
$w_2 \prec w_1 \approx w_2$
s_{13}
$w_1 \prec w_2 \approx w_3$
$w_1 \approx w_2 \approx w_3$
s_0

Figure B.3: *Structure $\langle S', \rightarrowtail', W', \preceq' \rangle$. The plausibility ordering of each stage is given in the corresponding box.*



(a) Topic lattice for A

(b) Topic lattice for B

Figure B.4: *Topic components for Contexts A & B. Topic assignment is given by labelling the nodes with atomic formulae.*

Tornadoes in Indonesia

As a last example, let's return to the case of the tornadoes in Indonesia presented in Section B.1.1. Consider the models $\mathcal{M}_A = \langle S', \rightarrowtail', W', \preceq', T', \oplus', t', \mathfrak{b}', \mathfrak{i}_A, V' \rangle$ and $\mathcal{M}_B = \langle S', \rightarrowtail', W', \preceq', T', \oplus', t', \mathfrak{b}', \mathfrak{i}_B, V' \rangle$, where $\langle S', \rightarrowtail', W', \preceq' \rangle$ is as in Figure B.3, with $V'(p) = \{w_1\}$ and $V'(q) = V'(r) = \{w_1, w_2\}$. The topic components $\langle T', \oplus', t', \mathfrak{b}', \mathfrak{i}_A \rangle$ and $\langle T', \oplus', t', \mathfrak{b}', \mathfrak{i}_B \rangle$ are as given in Figures B.4(a) and B.4(b), respectively. $\mathcal{M}_A$ and $\mathcal{M}_B$ are intended to model two distinct imaginative episodes of the same agent, where the distinction is solely due to the difference between the overall topics of the corresponding episodes. Thus, the only difference between the two models is the designated overall topics: $\mathfrak{i}_A$ and $\mathfrak{i}_B$. Now, let $p :=$ 'There is a tornado in Indonesia' be the explicit input, $q :=$ 'Booking a flight through the US rather than Indonesia is safer,' and $r :=$ 'My friend is in danger'. Then, $\mathcal{M}_A$ and $\mathcal{M}_B$ can be seen, respectively, as models of **Context A** and **Context B** from page 273. Suppose further that $\langle w_1, (s_0, s_{13}, s_{22}) \rangle$ is the actual world-history pair. We then have that $\mathcal{M}_A, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \Vdash Bq \wedge Br$ and $\mathcal{M}_B, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \Vdash Bq \wedge Br$. However, $\mathcal{M}_A, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \Vdash Iq$ (since $\preceq_{22} = \preceq_{13}^q$, $\mathcal{M}_A, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \Vdash Bq$, and $t'(q) \sqsubset \mathfrak{b}' \sqcap \mathfrak{i}_A$), but $\mathcal{M}_B, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \nVdash Iq$ (since $t'(q) \not\sqsubseteq \mathfrak{b}' \sqcap \mathfrak{i}_B$). Similarly, we also have $\mathcal{M}_A, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \nVdash Ir$ and $\mathcal{M}_B, \langle w_1, (s_0, s_{13}, s_{22}) \rangle \Vdash Ir$.

Appendix C

Metaphysics of Kinds

In this appendix, I will briefly discuss some of the main theories of the metaphysics of kinds, as mentioned in Chapter 8 (Section 8.2). This discussion is neither supposed to be exhaustive, nor definitive. I merely discuss some initial issues for some of these theories (see Khalidi (2013) for an excellent overview with an eye towards a Simple Causal Theory of kinds).

C.1 Conventionalism

As a metaphysical theory of kinds, *conventionalism* holds that "the differences and similarities that we attribute to things [i.e., kinds] exist in virtue of, for example, social function of the relevant concepts rather than in natural fact" (Bird & Tobin, 2018, §1.1.2). For example, the reason why we take the set of cats to be a natural kind, whereas we don't take the set of all white objects to be one is because there is a social, cognitive, or otherwise subjective relevance to the former set of objects that the latter set lacks. There is nothing objective, in nature, that makes it so that the one class is a natural kind and the other is not. In its strong form, conventionalism holds that different interests give rise to different kinds and none is more privileged than the other. This means that even the kinds that our best scientific practices focus on and uncover are not more objective than any of these other classifications. It all depends on human interests, not on what nature is really like.

The main issue with such a view is that it cannot explain why certain classifications of objects (i.e., kinds) are epistemically very fertile, whereas others (i.e., non-kinds) are not. Many agree that one of the main features of kindhood is that it allows for successful ampliative inferences (Quine, 1969; Hacking, 1991; Kornblith, 1993; Millikan, 2000; Khalidi, 2013; Bird & Tobin, 2018). E.g., by seeing a particular cat eat fish, it is reasonable to conclude that that cat will also eat fish at a later time or that other cats could eat fish. In a sense, if I see one cat eating fish, then that "knowledge will remain good on other encounters with cats" (Millikan, 2000, p. 3). However, when we think of random sets of objects, then there is no reason to think that our ampliative inferences will be successful. Take an example from Mallozzi (2018a, p. 14): "think of a bunch of random things that we decide belong to the same kind simply because they are all from New York: say, the Empire State Building, my super Joe, the Yankees, and the delicious everything bagels." If I know that everyday bagels are edible, I should not conclude that the Empire State Building is edible as well. We want a theory of kinds that is able to explain the difference in epistemic fertility between the set of cats and the set of things from New York.

Conventionalism is no such theory; it does not seem to be able to explain this difference. Relevance to our interest does not seem to bestow epistemic usefulness to sets of objects. It is not that *because* cats are more relevant to our interests that ampliative inferences with regards to them are more successful than inferences with respect to things in New York. Of course, it is the case that the properties and objects that we *focus* on depends, to some extent, on our interests, but what the conventionalists claim is something stronger, namely that our interests *determine* what kinds are.

C.2 Metaphysical Essentialism

Metaphysical essentialism, on the other hand, explicitly distinguishes between kinds and non-kinds.[1] It holds that what it is for an object to be a member of a kind is for it to have a particular essence, which other members of that kind also have. Objects in a random set, e.g., things from New York, do not share an essence. Metaphysical essentialism has it roots in Aristotle's metaphysics (Cohen, 2016) and results in a distinctive notion of *metaphysical* necessity. However, given the peculiarities of the Aristotelian metaphysics that give rise to this notion of essence (e.g., Aristotle's distinction between form and matter), most contemporary metaphysical essentialists base themselves on Kripke's (1980) arguments rather than Aristotle's (Priest, 2018).

The main characteristic of essentialism is that having the right essence is *necessary* and *sufficient* for being a member of the corresponding kind (see Khalidi, 2013, sec. 1.3 for a discussion on the different features of metaphysical essentialism with regards to kinds). There are two subtly different ways how one might interpret the modal implications of essentialism (see Khalidi, 2013, sec. 1.5). First of all, one might hold that being a natural kinds implies that its members belong to that kind in every world in which the members exist. That is, "[i]f i is a member of natural kind K in the actual world, then it is a member of that kind in every world in which i exists" (Khalidi, 2013, p. 22). This is a modal implication for the *members* of a kind: if you are a member of a particular kind in the actual world, then you are

[1] I will focus in this appendix on essentialism with regards to kinds. However, the notion of 'essence' generally applies more broadly than this. Examples are: origin essentialism (where a thing comes from), constitution essentialism (what a thing is made up of), et cetera.

a member of that kind in all worlds in which you exist. However, there is another modal implication of essentialism.

> [A] natural kind is one that is necessarily associated with a certain set of properties. That is, if a natural kind K is associated with properties $P_1, \ldots, P_n$, in the actual world, then K is associated with those very same properties in every possible world in which the kind is instantiated.
>
> (ibid.)

This second modal implication focuses on the *properties* that members of a kind necessarily have. Khalidi (2013, p. 23) argues that these modal implications are logically independent of each other, however, he notes that if the first modal implication is true *without* the second, it "rings rather hollow".

Even though metaphysical essentialism seems relatively popular among metaphysicians and philosophers of language, those working on natural kinds (and the psychology thereof) almost unanimously *reject* it. The reason being precisely these modal implications; to the first one there seem to be ample counterexamples and the second one seems unmotivated. Let us look at these objections in turn.

Consider again the first modal implication – i.e., that members belong to a kind necessarily. There are plenty counterexamples to this implication of essentialism from biological kinds (e.g., Dupré, 1981; Millikan, 2000; Khalidi, 2013; Bird & Tobin, 2018), so let us instead focus on a counterexample from a *physical* kind. The example involves a proton changing into an antiproton *in the actual world* and is worth quoting at length.

> When iridium nuclei are bombarded with protons, antiprotons are produced as a result. In this interaction, as it is typically described, protons are transformed into antiprotons. Hence, an individual proton may not remain a proton in the actual world. [...] From the fact that a proton *can become* an antiproton in the actual world, we might reasonably conclude that that proton *could have been* an antiproton in some other possible world. Thus, it seems that essentialists are wrong to insist that a proton could not have been anything but a proton in every possible world in which it exists. (Khalidi, 2013, p. 25, original emphases)

From examples like these, we should conclude that it is not at all obvious that members of a kind *necessarily* belong to that kind (Khalidi, 2013, p. 28).

Turn now to the second modal implication – i.e., that a kind is necessarily associated with a particular set of properties. It is this thesis, Khalidi argues, that is most closely associated with the kind-identities familiar from Putnam (1973) and Kripke (1980). Famous examples include water necessarily being H_2O and gold necessarily having atomic number 79. Kripke and, especially, Putnam argued for a particular *semantics* of natural kind terms, namely that they are rigid designators: they pick out the same kind in all possible worlds (in which that kind is instantiated). Based on this, many have interpreted Putnam as having shown that kinds have essences in the sense that there is a set of properties that members of that kind have by metaphysical necessity.[2] However, Salmon (1981) has forcefully argued that we cannot draw such metaphysical conclusions from a semantic theory without already sneaking in strong metaphysical assumptions (see also Mumford, 2005). For example, in Putnam's Twin Earth example, we hold fixed the chemical composition of water (i.e., H_2O) and conclude that, even though the watery-stuff on Twin Earth shares most other properties of water, Twin Earth watery-stuff *cannot be* water since it has a different chemical composition (e.g., XYZ). But why think that the chemical composition is what we should hold fixed? What if we hold fixed the *macroscopic* properties of water in the actual world (e.g., Chalmers, 1996; Jackson, 1998)? "[T]hen Putnam's conclusion that there is no water on Twin Earth and that water is necessarily H_2O does not follow from his thought experiment" (Khalidi, 2013, p. 28) (see also Priest, 2018, sec. 4.4).

Additionally, it is now generally accepted that kind terms are *not* rigid designators to begin with. It will be useful to go over a dilemma for the idea that kind terms refer to the same things in all possible worlds in order to suggest that the second modal implication of essentialism is generally unmotivated. The question is, what does one means with 'refer to the *same things*'?[3] An initial idea would be that 'same things' is meant to refer to all the actual individual members of a kind (past, present, and future). If this is what defenders of the rigidity of kind terms have in mind, then they are clearly wrong. Consider again the set of all cats and the, uncontroversial claim, that all of them could have a sibling more than they actually have. Now consider a world where the only cats are those possible siblings. As Khalidi notes, "[i]t seems uncontroversial to say that our term '[cat]' when applied to this possible world would pick out these individuals though none of them are identical to the actually existing individuals in this world" (2013, p. 29). So, it seems that kind terms *do not* pick out the actual members of that kind. On the other hand, if the defender of the rigidity of kind terms means that the terms pick out the same *kind* of things, then the demand on terms to count as rigid is trivial. That is, the demand is no longer something that distinguishes between natural and nonnatural kinds. To see this, consider again the set of things in New York, clearly a nonnatural kind, and use the term 'NY-things' to refer to it. On this interpretation of 'same things' the term 'NY-things' picks out the same kind of thing in every possible world, namely the kind of thing that is associated with the property of being in New York. This suggests that the "difference between natural kinds and nonnatural kinds is not that there is a *metaphysically* necessary connection between natural kinds and their associated properties, for the same could be said of nonnatural kinds" (Khalidi,

[2] Putnam (1990) later distanced himself from such a strict metaphysical interpretation.

[3] This paragraph is based on the discussion of Khalidi (2013, pp. 28-31).

2013, p. 30, emphasis added). But this was exactly what metaphysical essentialists suggested: that what makes natural kinds natural kinds is that they are necessarily associated with their essence.

There are ways in which an essentialist might respond and push back here – for example, they might suggest that we should only consider kinds with a single substance. However, as Khalidi points out, this would be a difference between natural kinds and nonnatural kinds related to the composition of their substance, *not* to the modal properties of natural kinds.

The above arguments count heavily against metaphysical essentialism with respect to kinds. Surely, the essentialist might hold that something 'weaker' than sets of necessary and sufficient properties are shared by all members of a kind and call this the 'essence'. However, this would not be metaphysical essentialism. As Millikan puts it,

> Most [natural] kinds do not have traditional essences [. . .] We could extend the term 'essence' so that it applies to whatever natural principle accounts for the instances of a kind being alike. But it is probably safer to [use another term] to avoid any possibility of misunderstanding.
> (2000, p. 23)

C.3 Homeostatic Property Cluster Theory

The homeostatic property cluster theory (HPC theory) is one of the most prominent theories of kinds amongst philosophers of science (Khalidi, 2013, p. 72). This theory follows the Quinean (1969) intuition that kind members share a number of important properties that are contingently (*pace* metaphysical essentialism) clustered together. In particular, HPC theorists hold that these properties cluster together due to an underlying causal mechanism that they take to be a *homeostatic mechanism* (Boyd, 1991, 1999). This homeostatic mechanism (or process) clusters together the shared properties of the kind members by keeping them in an equilibrium. As Bird and Tobin put it,

> Homeostatic property clusters occur when mechanisms exist that cause the properties to cluster by ensuring that deviations from the cluster have a low chance of persisting; the presence of some of the properties in the cluster favours the presence of the others. A homeostatic mechanism thereby achieves self-regulation, maintaining a stable range of properties.
> (2018, §1.2.2)

So, the homeostatic property cluster theory holds that what distinguishes natural kinds from nonnatural kinds is that the former have such a homeostatic mechanism to cluster the properties that they are associated with. It seems to have the benefit of metaphysical essentialism, accounting for the informativeness of certain classifications (i.e., the homeostatic process or mechanism), while not positing kinds to have a metaphysical essence with its problematic modal implications.

However, it has been noted that focusing on homeostatic processes or mechanisms does not seem suitable for *all* kinds of kinds. For example, proponents of the HPC theory themselves suggest that their theory is supposed to account for biological kinds rather than kinds from physics or chemistry. Others have argued that even within the case of biology, there are some kinds that fail to be classified as such on the HPC theory (see also Ereshefsky, 2010; Khalidi, 2013, pp. 74-75). For example, Ereshefsky argues that the theory is not compatible with the main theory of biological systematising, i.e., *cladistic* approaches, which focus on capturing historic descent (2010, p. 676).[4] Though HPC theorists might point out that there are *other* theories of systematisation available, Khalidi (2013, p. 75) points out that there is a more general worry in the vicinity of Ereshefsky's argument.

Remember the hierarchy thesis of kinds discussed in Chapter 8 (Section 8.2): if an object belongs to more than one kind, these kinds form a nested hierarchy. Especially in biology, we might want to say higher biological taxa (e.g., genera, families, etc.) are instances of such nested kinds. "If any organisms from different species are members of the same genus, then all members of both species are members of that genus" (Bird & Tobin, 2018, §1.1.1). However, accepting this seems to weaken the arguments in favour of HCP. The problem, as Khalidi (2013, p. 75) puts it, is that when it comes to higher biological taxa "the only serious candidate for a mechanism is genealogical descent. But if that is the case, then it might seem as though there is no work left to do for the homeostatic property cluster."

These are not devastating objections to the HPC theory of kinds, but combined with the fact that it seems to be unable to account for kinds in physics and chemistry, it is enough to warrant looking for an improvement on the theory.

C.4 Simple Causal Theory[5]

There is a natural suggestion to overcome these issues: "reject the [homeostatic processes and mechanisms] and keep the rest as a *simple causal theory* of natural kinds" (Craver, 2009, p. 579, original emphasis). This gets rid of the too stringent demand of the properties always being in equilibrium and, more importantly, allows us to explain more general relations between properties. The *Simple Causal Theory* does just that: natural kinds, as opposed to nonnatural kinds, have a set of properties

[4] Roughly, cladistic approaches try to maximise the fit of phylogenetic trees by counting the characteristics (or traits) that are possessed or absent within each of the taxa (see Bartha, 2010, pp. 202-203 for a toy example and Kitching et al., 1998 for a detailed discussion of cladistics).

[5] This part overlaps with Section 8.2.1 from Chapter 8.

(their 'causal core') that are causally related to a wide variety of other properties (and behaviours) often shared by members of that kind (Craver, 2009; Khalidi, 2013, 2018; Mallozzi, 2018a; Godman et al., 2020). In contrast to the HPC theory, the simple causal theory does not propose that this set of core properties is held together by any (homeostatic) mechanism.

This is very similar to the minimal theory we discussed in Chapter 8 (Section 8.2). Being a member of a kind, K, means having a set of core properties, C_K. These core properties are what Khalidi calls the primary properties and the cause, in a broad sense of the word, many of the other properties (and behaviours) that members of a kind share. For example, when we focus on the kind SILVER, we know that pieces of silver share many properties, e.g., melting point, boiling point, conductivity of sorts, colour, potential chemical combinations, et cetera. In the case of SILVER, it is the property of having atomic number 47 that causes members of the kind to have (and thus share) many of these other properties (see Mallozzi, 2018a, p. 9 for a detailed discussion of the silver example). Khalidi, one of the most prominent defenders of the simple causal theory, summarises the theory as follows.

> Crucially, [...], there is a causal link between properties, with one or a few of the properties being causally prior to the others. What characterizes natural kinds is that, even when one or a few properties are central to a kind, there are a number of other properties associated with that kind that are causally related to them. It is this network of properties that seems to distinguish natural kinds from non-natural kinds. The causal relations between the properties in the network ensure that natural kinds are projectible and play a central role in inductive inference. (2013, p. 204).

That is, very roughly, the simple causal theory of kinds (see Keil, 1995; Craver, 2009; Khalidi, 2013, 2018; Mallozzi, 2018a; and Godman et al., 2020). Importantly, on the simple causal theory of kinds, it needs to be discovered by science which properties make up a core of a kind. Additionally, the simple causal theory of kinds, just like the HPC theory, captures two of the main thoughts about kinds: "kinds have something to do with causation" and "that each natural kind is associated with a loose set or cluster of properties" (Khalidi, 2018, p. 1379). The difference between *essentialism* and HPC or the simple causal theory is that according to the latter two what makes a member of a kind a member of that kind is *contingent* and potentially in flux; whereas according to the essentialists the core properties (i.e., the essence) are necessary and sufficient conditions to belong to a kind.

Bibliography

Adler, J. (2017). Epistemological Problems of Testimony. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2017 ed.

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Functions for Contraction and Revision. *Journal of Symbolic Logic*, *50*, 510–530.

Alexander, J., & Weinberg, J. M. (2014). The "Unreliability" of Epistemic Intuitions. In E. Machery, & E. O'Neill (Eds.) *Current Controversies in Experimental Philosophy*, (pp. 128–145). New York, NY.: Routledge.

Alston, W. P. (1985). Concepts of Epistemic Justification. *The Monist*, *68*(1), 57–89.

——— (1993). *The Reliability of Sense Perception*. Ithaca, NY.: Cornell University Press.

——— (1999). Perceptual Knowledge. In J. Greco, & E. Sosa (Eds.) *The Blackwell Guide to Epistemology*, (pp. 223–242). Malden, MA.: Blackwell Publishing.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

Anderson, M. (2014). *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA.: MIT Press.

Angelucci, A., & Arcangeli, M. (2019). Introduction: New Perspectives on Philosophical Thought Experiments. *Topoi*, *38*, 763–768.

Anscombe, G. (1975). Causality and Determination. In E. Sosa (Ed.) *Causation and Conditionals*, (pp. 63–81). Oxford: Oxford University Press.

Arcangeli, M. (2018). *Supposition and the Imaginative Realm.* New York, NY.: Routledge.

Aronson, J. L., Harré, R., & Cornell Way, E. (1995). *Realism Rescued: How Scientific Progress is Possible*. Chigaco, IL.: Open Court.

Balcerak Jackson, M. (2016). On the Epistemic Value of Imagining, Supposing, and Conceiving. In A. Kind, & P. Kung (Eds.) *Knowledge Through Imagination*, (pp. 42–60). Oxford: Oxford University Press.

——— (2018). Justification by Imagination. In F. Macpherson, & F. Dorsch (Eds.) *Perceptual Imagination and Perceptual Memory*, (pp. 209–226). Oxford: Oxford University Press.

Baltag, A., & Smets, S. (2006). Dynamic Belief Revision over Multi-Agent Plausibility Models. In G. Bonanno, W. van der Hoek, & M. Wooldridge (Eds.) *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT2006)*, (pp. 11–24). University of Liverpool.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences*, *22*(4), 577–609.

——— (2008). Grounded Cognition. *Annual Review of Psychology*, *59*, 617–645.

——— (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289.

Bartha, P. (2010). *By Parallel Reasoning*. Oxford: Oxford University Press.

——— (2019). Analogy and analogical reasoning. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2019 ed.

Bealer, G. (1998). Intuition and the autonomy of philosophy. In M. R. DePaul, & W. Ramsey (Eds.) *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry*, (pp. 201–40). Lanham, MD.: Rowman & Littlefield.

——— (2000). A Theory of the A Priori: Intuition, Evidence, Concept-Possession. *Pacific Philosophical Quarterly*, *81*(1), 1–30.

——— (2002). Modal Epistemology and the Rational Renaissance. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 71–126). Oxford: Oxford University Press.

Beebe, J. R., & Shea, J. (2013). Gettierized Knobe Effects. *Episteme*, *10*, 219–240.

Beebee, H. (2006). Does Anything Hold the Universe Together? *Synthese*, *149*(3), 509–533.

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Benacerraf, P. (1973). Mathematical truth. *The Journal of Philosophy*, *70*(19), 661–679.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

van Benthem, J. (2007). Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, *17*(2), 129–155.

Berto, F. (2013). *Existence as a Real Property. The Ontology of Meinongianism*, vol. 356 of *Synthese Library*. Dordrecht: Springer.

——— (2015). A modality called ‘negation’. *Mind*, *124*(495), 761–793.

——— (2017). Impossible Worlds and the Logic of Imagination. *Erkenntnis*, *82*(6), 1277–1297.

——— (2018a). Aboutness in imagination. *Philosophical Studies*, *175*, 1871–1886.

——— (2018b). Taming the runabout imagination ticket. *Synthese*, (forthcoming), 1–15. https://doi.org/10.1007/s11229-018-1751-6.

——— (2019). Simple hyperintensional belief revision. *Erkenntnis*, *84*, 559–575.

Berto, F., French, R., Priest, G., & Ripley, D. (2018). Williamson on Counterpossibles. *Journal of Philosophical Logic*, *47*, 693–713.

Berto, F., & Hawke, P. (2018). Knowability relative to information. *Mind*, (forthcoming), 1–33. https://doi.org/10.1093/mind/fzy045.

Berto, F., & Özgün, A. (2020). Dynamic Hyperintensional Belief Revision. *Review of Symbolic Logic*, (forthcoming), 1–46. https://doi.org/10.1017/S1755020319000686.

Berto, F., & Restall, G. (2019). Negation on the Australian Plan. *Journal of Philosophical Logic*, *48*, 1119–1144.

Berto, F., & Schoonen, T. (2018). Conceivability and possibility: some dilemmas for Humeans. *Synthese*, *195*(6), 2697–2715.

Biggs, S. (2011). Abduction and modality. *Philosophy and Phenomenological Research*, *83*(2), 283–326.

Binet, A. (1899). *The Psychology of Reasoning*. London: The Open Court Publishing Company. (A. G. Whyte, Trans.).

Bird, A. (2018). The metaphysics of natural kinds. *Synthese*, *195*, 1397–1426.

Bird, A., & Tobin, E. (2018). Natural kinds. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2018 ed.

Blackburn, S. (1993). Morals and Modals. In *Essays in Quasi-Realism*, (pp. 52–74). New York, NY.: Oxford University Press.

Bliss, R., & Trogdon, K. (2016). Metaphysical grounding. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2016 ed.

Bonanno, G. (2007). Axiomatic characterization of the AGM theory of belief revision in a temporal logic. *Artificial Intelligence*, *171*(2), 144 – 160.

——— (2012). Belief Change in Branching Time: AMG-consistency and Iterated Revision. *Journal of Philosophical Logic*, *41*(1), 201–236.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge, MA.: Harvard University Press.

——— (1998). *In defense of pure reason: A rationalist account of a priori justification*. Cambridge: Cambridge University Press.

——— (2003). A Version of Internalist Foundationalism. In L. BonJour, & E. Sosa (Eds.) *Epistemic Justification*, (pp. 3–96). Malden, MA.: Blackwell.

BonJour, L., & Sosa, E. (2003). *Epistemic Justification*. Malden, MA.: Blackwell Publishing.

Boyd, R. (1991). Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies*, *61*, 127–148.

——— (1999). Homeostasis, species, and higher taxa. In R. Wilson (Ed.) *Species: New interdisciplanary essays*. Cambridge, MA.: MIT Press.

Brown, J. (2018). *Fallibilism: Evidence and Knowledge*. Oxford: Oxford University Press.

Brown, J. R. (1986). Thought Experiments since the Scientific Revolution. *International Studies in the Philosophy of Science*, *1*, 1–15.

Brown, J. R., & Fehige, Y. (2019). Thought Experiments. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2019 ed.

Brueckner, A. (2001). Chalmers's Conceivability Argument for Dualism. *Analysis*, *61*(3), 187–193.

Bueno, O., & Shalkowski, S. (2014). Modalism and theoretical virtues: Toward an epistemology of modality. *Philosophical Studies*, *172*(3), 671–689.

Burr, C., & Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, *29*(4), 586–600.

Byrne, A. (2007). Possibility and Imagination. *Philosophical Perspectives*, *21*(1), 125–144.

Byrne, R. M. J. (2005). *The Rational Imagination*. London: The MIT Press.

——— (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(7), 7.1–7.23.

Cameron, R. P. (2010). Response to Dominic Gregory. In B. Hale, & A. Hoffmann (Eds.) *Modality. Metaphysics, Logic, and Epistemology*, (pp. 342–345). Oxford: Oxford University Press.

Canavotto, I., Berto, F., & Giordani, A. (2020). Voluntary Imagination: A Fine-Grained Analysis. *Review of Symbolic Logic*, (forthcoming), 1–34. https://doi.org/10.1017/S1755020320000039.

Carey, S. (1985). *Conceptual development in childhood*. Cambridge, MA.: MIT Press.

——— (2009). *The Origin of Concepts*. Oxford: Oxford University Press.

Carroll, J. W. (2009). Anti-Reductionism. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.) *The Oxford Handbook of Causation*, (pp. 279–298). Oxford: Oxford University Press.

Carroll, J. W., & Markosian, N. (2010). *An Introduction to Metaphysics*. Cambridge: Cambridge University Press.

Carruthers, P. (1996). Simulation and self-knowledge: A defense of theory-theory. In P. Carruthers, & P. Smith (Eds.) *Theories of Theories of Mind*, (pp. 22–68). Cambridge: Cambridge University Press.

Carruthers, P., & Smith, P. (Eds.) (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Casullo, A. (2010). Knowledge and modality. *Synthese*, *172*, 341–359.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

——— (1999). Materialism and the Metaphysics of Modality. *Philosophy and Phenomenological Research*, *59*, 473–493.

——— (2002). Does Conceivability Entail Possibility? In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 145–200). Oxford: Oxford University Press.

——— (2010). *The Character of Consciousness*. Oxford: Oxford University Press.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(3), 185–211.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA.: MIT Press.

Christensen, D. (2013). Epistemic Modesty Defended. In D. Christensen, & J. Lackey (Eds.) *The Epistemology of Disagreement*, (pp. 77–97). Oxford: Oxford University Press.

Cimpian, A. (2015). The Inherence Heuristic: Generating Everyday Explanations. In R. Scott, & S. Kosslyn (Eds.) *Emerging Trends in the Social and Behavioral Sciences*, (pp. 1–15). Wiley Online Library. https://doi.org/10.1002/9781118900772.etrds0341.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*, 461–480.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*. Cambridge, MA.: MIT press.

——— (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

——— (2016). *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

——— (2017). Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil. *Noûs*, *51*(4), 727–753.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*, 7–19.

Clarke-Doane, J. (2019a). Metaphysical and absolute possibility. *Synthese*, (forthcoming), 1–12. https://doi.org/10.1007/s11229-019-02093-0.

——— (2019b). Modal Objectivity. *Noûs*, *53*(2), 266–295.

Clavel-Vázquez, A., & Clavel Vázquez, M. J. (2018). Embodied Imagination: Why we can't just walk in someone else's shoes. Blogpost The Junkyard. https://junkyardofthemind.com/blog/2018/8/5/embodied-imagination-why-we-cant-just-walk-in-someone-elses-shoes (accessed July 18th 2020).

Cohen, H., & Lefebvre, C. (2005). Bridging the Category Divide. In H. Cohen, & C. Lefebvre (Eds.) *Handbook of Categorization in Cognitive Science*, (pp. 2–15). Oxford: Elsevier.

Cohen, J. (1988a). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ.: Lawrence Erlbaum Associates.

——— (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cohen, S. (1988b). How to be a fallibilist. In J. Tomberlin (Ed.) *Philosophical Perspective 2: Epistemology*, (pp. 91–123). Atascadero, CA.: Ridgeview Publishing Co.

Cohen, S. M. (2016). Aristotle's Metaphysics. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2016 ed.

Cohnitz, D., & Häggqvist, S. (2018). Thought Experiments in Current Metaphilosophical Debates. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.) *The Routledge Companion to Thought Experiments*, (pp. 406–424). New York, NY.: Routledge.

Coley, J. D., Medin, D. L., Proffitt, J. B., Lynch, E., & Atran, S. (1999). Inductive Reasoning in Folkbiological Thought. In D. L. Medin, & S. Atran (Eds.) *Folkbiology*, (pp. 205–232). Cambridge, MA.: MIT Press.

Collins, J. D., Hall, N., & Paul, L. (Eds.) (2004). *Causation and Counterfactuals*. Cambridge, MA.: MIT Press.

Comesaña, J., & Klein, P. (2019). Skepticism. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2019 ed.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, *22*(5), 575–594.

Currie, G., & Ravenscroft, I. (2002). *Recreative Minds*. Oxford: Oxford University Press.

Dancy, J., Sosa, E., & Steup, M. (Eds.) (2010). *A Companion to Epistemology*. Malden, MA.: Wiley-Blackwell.

Davidson, D. (1987). Knowing One's Own Mind. In *Proceedings and Addresses of the American Philosophical Association*, vol. 60, (pp. 441–458).

Davies, T. R. (1988). Determination, Uniformity, and Relevance: Normative Criteria for Generalization and Reasoning by Analogy. In D. H. Helman (Ed.) *Analogical Reasoning*, (pp. 227–250). Dordrecht: Kluwer Academic Publishers.

De, M., & Omori, H. (2018). There is more to negation than modality. *Journal of Philosophical Logic*, *47*, 281–299.

De Brigard, F., Addis, D., Ford, J., Schacter, D., & Giovanello, K. (2013). Remembering what could have happened: Neural correlates of episodic counterfactual thinking. *Neuropsychologia*, *51*(12), 2401–2414.

Deacon, T. W. (1997). *The Symbold Species: The co-evolution of language and the brain*. New York, NY.: W.W. Norton & Company, Inc.

Deák, G. O., & Bauer, P. J. (1996). The Dynamics of Preschoolers' Categorization Choices. *Child Development*, *67*(3), 740–767.

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, *29*(1), 1–22.

Demey, L. (2013). Contemporary Epistemic Logic and the Lockean Thesis. *Foundations of science*, *18*(4), 599–610.

Divers, J. (2002). *Possible Worlds*. London: Routledge.

Doggett, T., & Stoljar, D. (2010). Does Nagel's footnote eleven solve the mind-body problem? *Philosophical Issues*, *20*(1), 125–143.

Dohrn, D. (2010). Hume on knowledge of metaphysical modality. In H. Beebee, & M. Schrenk (Eds.) *David Hume: Epistemology and Metaphysics*, vol. 13, (pp. 38–59). Paderborn: Mentis.

——— (2019). Modal epistemology made concrete. *Philosophical Studies*, *176*, 2455–2475.

Dougherty, T. (2011). Fallibilism. In S. Bernecker, & D. Pritchard (Eds.) *The Routledge Companion to Epistemology*, (pp. 131–143). New York, NY: Routledge.

Douven, I. (2013). The Epistemology of Conditionals. In T. S. Gendler, & J. Hawthorne (Eds.) *Oxford Studies in Epistemology*, vol. 4, (pp. 3–33). Oxford: Oxford University Press.

——— (2015). *The epistemology of indicative conditionals: Formal and empirical approaches*. Cambridge: Cambridge University Press.

Douven, I., & Verbrugge, S. (2010). The Adams family. *Cognition*, *117*(3), 302–318.

——— (2013). The probabilities of conditionals revisited. *Cognitive Science*, *37*(4), 711–730.

Dove, G. (2014). Thinking in words: Language as an embodied medium of thought. *Topics in Cognitive Science*, *6*(3), 371–389.

Dretske, F. (1970). Epistemic Operators. *The Journal of Philosophy*, *67*(24), 1007–1023.

Driver, J. (1999). Modesty and Ignorance. *Ethics*, *109*, 827–834.

Dummett, M. (1959). Wittgenstein's Philosophy of Mathematics. *The Philosophical Review*, *68*(3), 324–348.

Dupré, J. (1981). Natural Kinds and Biological Taxa. *Philosophical Review*, *90*, 66–90.

Dutant, J. (2015). The legend of the justified true belief analysis. *Philosophical Perspectives*, *29*(1), 95–145.

Edgington, D. (1986). Do conditionals have truth conditions? *Crítica: Revista Hispanoamericana de Filosofía*, (pp. 3–39).

——— (1995). On conditionals. *Mind*, *104*(414), 235–329.

Elqayam, S., & Over, D. E. (2013). New paradigm psychology of reasoning. *Thinking & Reasoning*, *19*(3-4), 249–265.

Engel, P. (1998). Believing, holding true, and accepting. *Philosophical explorations*, *1*(2), 140–151.

Epstude, K., & Roese, N. J. (2008). The Functional Theory of Counterfactual Thinking. *Personality and Social Psychology Review*, *12*(2), 168–192.

Ereshefsky, M. (2010). What's wrong with the new biological essentialism. *Philosophy of Science*, *77*(5), 674–685.

Evnine, S. (2008). Modal epistemology: Our knowledge of necessity and possibility. *Philosophy Compass*, *3*(4), 664–684.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1990). The Struture-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, *41*, 1–63.

Fehige, Y. J. H., & Wiltsche, H. (2013). The Body, Thought Experiments, and Phenomenology. In M. Frappier, L. Meynell, & J. R. Brown (Eds.) *Thought Experiments in Science, Philosophy, and the Arts*, (pp. 69–89). New York, NY.: Routledge.

Feldman, R. (2014). Justification Is Internal. In M. Steup, J. Turri, & E. Sosa (Eds.) *Contemporary Debates in Epistemology*, (pp. 337–350). Malden, MA.: Wiley-Blackwell.

Fine, K. (1986). Analytic implication. *Notre Dame J. Formal Logic*, *27*(2), 169–179.

——— (1994). Essence and Modality: The Second Philosophilical Perspectives Lecture. *Philosophical Perspectives*, *8*, 1–16.

——— (2002). The varieties of necessity. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 253–281). Oxford: Oxford University Press.

——— (2016). Angellic content. *Journal of Philosophical Logic*, *45*(2), 199–226.

Fiocco, M. O. (2007). Conceivability, Imagination and Modal Knowledge. *Philosophy and Phenomenological Research*, *74*(2), 364–380.

Fischer, B. (2016a). Hale on the architecture of modal knowledge. *Analytic Philosophy*, *57*(1), 76–89.

——— (2016b). A theory-based epistemology of modality. *Canadian Journal of Philosophy*, *46*(2), 228–247.

——— (2017a). Modal empiricism: Objection, reply, proposal. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, (pp. 263–280). Cham: Springer.

——— (2017b). *Modal Justification via Theories*, vol. 380 of *Synthese Library. Studies in Epistemology, Logic, Methodology, and Philosophy of Science*. Cham: Springer.

Fischer, B., & Leon, F. (2017a). Introduction to *Modal Epistemology After Rationalism*. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, (pp. 1–6). Cham: Springer.

Fischer, B., & Leon, F. (Eds.) (2017b). *Modal Epistemology After Rationalism*, vol. 378 of *Synthese Library. Studies in Epistemology, Logic, Methodology, and Philosophy of Science*. Cham: Springer International Publishing.

Fitting, M., & Mendelsohn, R. L. (1998). *First-Order Modal Logic*. Synthese Library. Dordrecht: Kluwer Academic Publishers.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA.: Harvard University Press.

——— (1981). *Representations*. Cambridge, MA.: MIT Press.

Foley, R. (1992). The Epistemology of Belief and the Epistemology of Degrees of Belief. *American Philosophical Quarterly*, *29*(2), 111–124.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review*, *5*, 5–15.

Forbus, K. D., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998). Analogy just looks like high level perception: Why a domain-general approach to analogical mapping is right. *Journal of Experimental & Theoretical Artificial Intelligence*, *10*(2), 231–257.

Fricker, E. (2016). Unreliable Testimony. In B. P. McLaughlin, & H. Kornblith (Eds.) *Goldman and His Critics*, (pp. 88–120). Malden, MA.: Wiley-Blackwell.

Fumerton, R. (2002). Theories of Justification. In P. K. Moser (Ed.) *The Oxford Handbook of Epistemology*, (pp. 204–233). Oxford: Oxford University Press.

Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 659–669.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, *2*(12), 493–501.

Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, MA.: MIT Press.

Geddes, A. (2017). Judgements about Thought Experiments. *Mind*, *127*(505), 35–67.

Geirsson, H. (2005). Conceivability and Defeasible Modal Justification. *Philosophical Studies*, *122*(3), 279–304.

Gelman, S. A. (2003). *The Essential Child: Origins of essentialism in everyday thought*. Oxford: Oxford University Press.

——— (2019). What the Study of Psychological Essentialism May Reveal about the Natural World. In A. I. Goldman, & B. P. McLaughlin (Eds.) *Metaphysics and Cognitive Science*, (pp. 314–333). New York, NY.: Oxford University Press.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, *26*, 796–804.

——— (1991). The acquisition of natural kind terms. In S. A. Gelman, & J. P. Byrnes (Eds.) *Perspectives on language and thought: Interrelations in development*, (pp. 146–196). Cambridge: Cambridge University Press.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

——— (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532–1541.

Gelman, S. A., & Meyer, M. (2011). Child categorization. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 95–105.

Gendler, T. S. (2000). The puzzle of imaginative resistance. *Journal of Philosophy*, *97*(2), 55–81.

Gendler, T. S., & Hawthorne, J. (Eds.) (2002a). *Conceivability and Possibility*. Oxford: Oxford University Press.

Gendler, T. S., & Hawthorne, J. (2002b). Introduction. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 1–70). Oxford: Oxford University Press.

——— (2005). The real guide to fake barns: A catalogue of gifts for your epistemic enemies. *Philosophical Studies*, *124*, 331–352.

Gentner, D. (1983). Structure Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, *7*, 155–170.

——— (1989). The mechanisms of analogical learning. In S. Vosniadou, & A. Ortony (Eds.) *Similarity and Analogical Reasoning*, (pp. 199–241). Cambridge: Cambridge University Press.

Gentner, D., & Jeziorski, M. (1993). The shift from metaphor to analogy in Western science. In A. Ortony (Ed.) *Metaphor and Thought*, (pp. 447–480). Cambridge: Cambridge University Press.

Gentner, D., & Markman, A. B. (1997). Structure Mapping in Analogy and Similarity. *American Psychologist*, *52*(1), 45–56.

Gettier, E. (1963). Is Justified True Belief Knowledge? *Analysis*, *23*, 121–123.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. New York, NY.: Taylor & Francis. Psychology Press Classic Edition (2015).

Gigerenzer, G. (1996). On narrow norms and vague heuristics: a reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

——— (1998). Ecological intelligence: an adaptation for frequencies. In D. Cummins, & C. Allen (Eds.) *The evolution of mind*, (pp. 9–29). New York, NY.: Oxford University Press.

Gigerenzer, G., & Murrya, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ.: Erlbaum.

Godman, M., Mallozzi, A., & Papineau, D. (2020). Essential Properties are Super-Explanatory: Taming Metaphysical Modality. *Journal of American Philosophical Association*, (forthcoming), 1–19. https://doi.org/10.1017/apa.2019.48.

Goldberg, S. C. (2015). What Is the Subject-matter of the Theory of Epistemic Justification? In D. K. Henderson, & J. Greco (Eds.) *Epistemic Evaluation: Purposeful Epistmeology*, (pp. 205–223). Oxford: Oxford University Press.

Goldman, A. I. (1976). Discrimination and perceptual knowledge. *The Journal of Philosophy*, *73*, 771–791.

——— (1979). What Is Justified Belief? In G. S. Pappas (Ed.) *Justification and Knowledge. New Studies in Epistemology*, (pp. 1–23). Dordrecht: D. Reidel.

——— (1992). *Liaisons: philosophy meets the cognitive and social sciences*. Malden, MA.: MIT Press.

——— (1994). Naturalistic epistemology and reliabilism. *Midwest studies in philosophy*, *19*, 301–320.

——— (1999). Internalism exposed. *The Journal of Philosophy*, *96*(6), 271–293.

——— (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford: Oxford University Press.

——— (2012). *Reliabilism and Contemporary Epistemology*. Oxford: Oxford University Press.

Goldston, D. B., Hinrichs, J. V., & Richman, C. L. (1985). Subjects' expectations, individual variability, and the scanning of mental images. *Memory & Cognition*, *13*(4), 365–370.

Gooding, D. (1992). What is experimental about thought experiments? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1992*(2), 280–290.

——— (1994). Imaginary science. *The British Journal for the Philosophy of Science*, *45*(4), 1029–1045.

Goodman, N. (1972). Seven Strictures on Similarity. In *Problems and Projects*, (pp. 437–446). Indianapolis, IN.: Bobbs-Merrill.

——— (1976). *Languages of Art*. Indianapolis, IN.: Hackett Publishing Company.

Gopnik, A., & Walker, C. M. (2013). Considering Counterfactuals. The Relationship between Causal Learning and Pretend Play. *American Journal of Play*, *6*(1), 15–28.

Greco, J. (2014). Justification Is Not Internal. In M. Steup, J. Turri, & E. Sosa (Eds.) *Contemporary Debates in Epistemology*, (pp. 325–336). Malden, MA.: Wiley-Blackwell.

Gregory, D. (2004). Imagining Possibilities. *Philosophy and Phenomenological Research*, *69*(2), 327–348.

——— (2010). Conceivability and Apparent Possibility. In B. Hale, & A. Hoffmann (Eds.) *Modality. Metaphysics, Logic, and Epistemology*, (pp. 319–341). Oxford: Oxford University Press.

——— (2017). Counterfactual reasoning and knowledge of possibilities. *Philosophical Studies*, *174*(4), 821–835.

——— (2019). Imagery and Possibility. *Noûs*, (forthcoming), 1–19. https://doi.org/10.1111/nous.12275.

Grundmann, T., & Horvath, J. (2014). Thought experiments and the problem of deviant realizations. *Philosophical Studies*, *170*(3), 525–533.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*, 377–397.

Hacking, I. (1991). A tradition of natural kinds. *Philosophical Studies*, *61*(1), 109–126.

Hale, B. (2003). The Presidential Address: Knowledge of Possibility and of Necessity. *Proceedings of the Aristotelian Society*, *103*(1), 1–20.

——— (2013). *Necessary beings: An essay on ontology, modality, and the relations between them*. Oxford: Oxford University Press.

Hanrahan, R. (2017). Imagination, Possibility, and Plovers. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, (pp. 197–219). Cham: Springer.

Harnad, S. (2005). To Cognize is to Categorize: Cognition is Categorization. In H. Cohen, & C. Lefebvre (Eds.) *Handbook of Categorization in Cognitive Science*, (pp. 20–43). Oxford: Elsevier.

Harris, P. L. (2000). *The Work of the Imagination.* Malden, MA.: Blackwell.

Hartl, P. (2016). Modal scepticism, Yablo-style conceivability, and analogical reasoning. *Synthese*, *193*(1), 269–291.

Hawke, P. (2011). Van Inwagen's modal skepticism. *Philosophical Studies*, *153*(3), 351–364.

——— (2017). Can Modal Skepticism Defeat Humean Skepticism? In B. Fischer, & F. Leon (Eds.) *Modal Epistemology after Rationalism*, vol. 378, (pp. 281–308). Cham: Springer.

——— (2018). Theories of aboutness. *Australasian Journal of Philosophy*, *96*(4), 697–723.

Hawke, P., Berto, F., & Özgün, A. (2020). The Fundamental Problem of Logical Omniscience. *Journal of Philosophical Logic*, *49*, 727–766.

Hawke, P., & Schoonen, T. (2020). Are Gettier Cases Disturbing? *Philosophical Studies*, (forthcoming), 1–25. https://doi.org/10.1007/s11098-020-01493-0.

Hawley, K., & Bird, A. (2011). What are natural kinds? *Philosophical Perspectives*, *25*, 205–221.

Hawthorne, J. (2004). *Knowledge and Lotteries.* Oxford: Oxford University Press.

——— (2009). The Lockean Thesis and the Logic of Belief. In F. Huber, & C. Schmidt-Petri (Eds.) *Degrees of belief*, (pp. 49–74). Dordrect: Springer.

Hayes, B. K., & Thompson, S. P. (2007). Causal relations and feature similarity in children's inductive reasoning. *Journal of Experimental Psychology: General*, *136*(3), 470.

Helman, D. H. (Ed.) (1988). *Analogical Reasoning. Perspectives of Artificial Intelligence, Cognitive Science, and Philosophy*, vol. 197 of *Synthese Library*. Dordrecht: Kluwer Academic Publishers.

Hesse, M. B. (1966). *Models and Analogies in Science.* Notre Dame, IN: University of Notre Dame Press.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences*, *6*(6), 242–247.

Hill, C. S. (1997). Imaginability, Conceivability, Possibility and the Mind-Body Problem. *Philosophical Studies*, *87*, 61–85.

——— (2006). Modality, Modal Epistemology, and the Metaphysics of Consciousness. In S. Nichols (Ed.) *The Architecture of the Imagination*, (pp. 205–236). Oxford: Oxford University Press.

——— (2016). Conceivability and Possibility. In H. Cappelen, T. S. Gendler, & J. Hawthorne (Eds.) *The Oxford Handbook of Philosophical Methodology*, (pp. 326–347). Oxford: Oxford University Press.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, NY.: Cornell University Press.

Holden, T. (2014). Hume's Absolute Necessity. *Mind*, *123*(490), 377–413.

Holliday, W., & Icard, T. F. (2010). Moorean phenomena in epistemic logic. In L. Beklemishev, V. Goranko, & V. Shehtman (Eds.) *Advances in Modal Logic*, vol. 8, (pp. 178–199). London: College Publications.

Hume, D. (1777/1997). *An Enquiry Concerning Human Understanding*. Indianapolis, IN.: Hackett Publishing Company.

Humphreys, P. (1980). Cutting the Causal Chain. *Pacific Philosophical Quarterly*, *61*, 305–314.

——— (1981). Aleatory Explanations. *Synthese*, *48*, 225–232.

Hutto, D., & Myin, E. (2012). *Radicalizing encativism: Basic minds without content*. Cambridge, MA.: MIT Press.

Hutto, D. D. (2015). Overly Enactive Imagination? Radically Re-Imagining Imagining. *The Southern Journal of Philosophy*, *53*(Spindel Supplement), 68–89.

Ichikawa, J. J. (2009). Knowing the Intuition and Knowning the Counterfactual. *Philosophical Studies*, *145*(3), 435–443.

Ichikawa, J. J., & Jarvis, B. W. (2009). Thought-Experiment Intuitions and Truth in Fiction. *Philosophical Studies*, *142*(2), 221–246.

——— (2012). Rational Imagination and Modal Knowledge. *Noûs*, *46*(1), 127–158.

——— (2013). *The Rules of Thought*. Oxford: Oxford University Press.

Intons-Peterson, M. J. (1983). Imagery Pardigms: How Vulnerable Are They to Experimenter's Expectations? *Journal of Experimental Psychology: Human Perception and Performance*, *9*(3), 394–412.

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, *83*(5), 291–295.

——— (1998). *From Metaphysics to Ethics. A Defence of Conceptual Analysis*. Oxford: Clarendon Press.

Jago, M. (2014). *The Impossible*. Oxford: Oxford University Press.

——— (2018). Knowing how things might have been. *Synthese*, (forthcoming), 1–19. https://doi.org/10.1007/s11229-018-1869-6.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences*, *17*(2), 187–202.

——— (2006). *Motor Cognition*. Oxford: Oxford University Press.

Jenkins, C. S. (2008). Modal Knowledge, Counterfactual Knowledge and the Role of Experience. *The Philosophical Quarterly*, *58*(233), 693–701.

——— (2013). Naturalistic Challenges to the A Priori. In A. Casullo, & J. C. Thurow (Eds.) *The A Priori in Philosophy*, (pp. 274–290). Oxford: Oxford University Press.

Jones, M. (2018). Seeing numbers as affordances. In S. Bangu (Ed.) *Naturalizing Logic-Mathematical Knowledge*, (pp. 148–163). New York, NY.: Routledge.

Jones, M., & Wilkinson, S. (2020). From Prediction to Imagination. In A. Abraham (Ed.) *The Cambridge Handbook of the Imagination*, (pp. 94–110). Cambridge: Cambridge University Press.

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY.: Farrar, Straus and Giroux.

Kail, P. (2003). Conceivability and modality in hume: A lemma in an argument in defense of skeptical realism. *Hume Studies*, *29*(1), 43–61.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA.: The MIT Press.

——— (1995). The growth of causal understandings of natural kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.) *Causal Cogntion: A multidisciplinary debate*, (pp. 234–267). Oxford: Oxford University Press.

Khalidi, M. A. (1993). Carving Nature at the Joints. *Philosophy of Science*, *1993*(1), 100–113.

——— (1998). Natural Kinds and Crosscutting Categories. *The Journal of Philosophy*, *95*(1), 33–50.

——— (2013). *Natural categories and human kinds: Classification in the natural and social sciences*. Cambridge: Cambridge University Press.

——— (2018). Natural kinds as nodes in causal networks. *Synthese*, *195*(4), 1379–1396.

Kim, H., Kneer, M., & Stuart, M. T. (2019). The Content-Dependence of Imaginative Resistance. In F. Cova, & S. Réhault (Eds.) *Advances in Experimental Philosophy of Aesthetics*, (pp. 143–166). London: Bloomsbury Academic.

Kim, M., & Yuan, Y. (2015). No Cross-cultural Differences in Gettier Case Case Intuition: A Replication Study of Weinberg et al. 2001. *Episteme*, *12*(3), 355–361.

Kind, A. (2001). Putting the image back in imagination. *Philosophy and Phenomenological Research*, *62*(1), 85–109.

——— (2013). The heterogeneity of the imagination. *Erkenntnis*, *78*(1), 141–159.

——— (2016a). Imagining Under Constraints. In A. Kind, & P. Kung (Eds.) *Knowledge Through Imagination*, (pp. 145–159). Oxford: Oxford University Press.

——— (2016b). Introduction: Exploring imagination. In A. Kind (Ed.) *The Routledge Handbook of Philosophy of Imagination*, (pp. 1–11). London: Routledge.

Kind, A. (Ed.) (2016c). *The Routledge Handbook of Philosophy of Imagination*. London: Routledge.

Kind, A., & Kung, P. (2016a). Introduction: The Puzzle of Imaginative Use. In A. Kind, & P. Kung (Eds.) *Knowledge Through Imagination*, (pp. 1–37). Oxford: Oxford University Press.

Kind, A., & Kung, P. (Eds.) (2016b). *Knowledge Through Imagination*. Oxford: Oxford University Press.

Kirchhoff, M. D. (2018). Preditive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philosophical Studies*, *175*, 751–767.

Kitching, I., Forey, P., Humphries, C., & Williams, D. (1998). *Cladistics: The Theory and Practice of Parsimony Analysis*. Oxford: Oxford University Press, 2nd ed.

Kment, B. (2006). Counterfactuals and the Analysis of Necessity. *Philosophical Perspectives*, *20*, 237–302.

——— (2014). *Modality and explanatory reasoning*. Oxford: Oxford University Press.

——— (2017). Varieties of modality. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2017 ed.

——— (2018). Essence and modal knowledge. *Synthese*, (forthcoming), 1–23. https://doi.org/10.1007/s11229-018-01903-1.

Kornblith, H. (1993). *Inductive Inference and Its Natural Ground. An Essay in Naturalistic Epistemology*. Cambridge, MA.: The MIT Press.

Kornblith, H. (Ed.) (2001). *Epistemology: Internalism and Externalism*. Malden, MA.: Blackwell Publishers.

Kosslyn, S. M. (1973). Scanning visual images: Some structural implications. *Perception & Psychophysics*, *14*(1), 90–94.

——— (1980). *Image and Mind*. Cambridge, MA.: Harvard University Press.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of experimental psychology: Human perception and performance*, *4*(1), 47–60.

Kripke, S. (1971). Identity and Necessity. In M. K. Munitz (Ed.) *Identity and Individuation*, (pp. 135–164). New York, NY.: New York University Press.

——— (1980). *Naming and Necessity*. Oxford: Blackwell Publishers.

Kroedel, T. (2012). Counterfactuals and the Epistemology of Modality. *Philosophers' Imprint*, *12*(12), 1–14.

——— (2017). Modal Knowledge, Evolution, and Counterfactuals. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 179–195). Cham: Springer.

Kung, P. (2010). Imagining as a Guide to Possibility. *Philosophy and Phenomenological Research*, *81*(3), 620–663.

——— (2016). You Really Do Imagine It: Against Error Theories of Imagination. *Noûs*, *50*(1), 90–120.

——— (2017). Personal Identity Without Too Much Science Fiction. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 133–154). Cham: Springer International Publishing.

Lam, D. (2017). Is imagination too liberal for modal epistemology? *Synthese*, *195*(5), 2155–2174.

Lammenranta, M. (2004). Theories of Justification. In I. Niiniluoto, M. Sintonen, & J. Woleński (Eds.) *Handbook of Epistemology*, (pp. 467–497). Dordrecht: Kluwer Academic Publishers.

Lane, J. D., Ronfard, S., Francioli, S., & Harris, P. L. (2016). Children's imagination and belief: Prone to flights of fancy or grounded in reality? *Cognition*, *152*, 127–140.

Langland-Hassan, P. (2012). Pretense, imagination, and belief: the Single Attitude theory. *Philosophical Studies*, *159*, 155–179.

——— (2016). On Choosing What to Imagine. In A. Kind, & P. Kung (Eds.) *Knowledge Through Imaginaion*, (pp. 61–84). Oxford: Oxford University Press.

Laskowski, N. (2018). Epistemic modesty in ethics. *Philosophical Studies*, *175*, 1577–1596.

Leahy, B. P., & Carey, S. E. (2020). The Acquisition of Modal Concepts. *Trends in Cognitive Sciences*, *24*(1), 65–78.

Leftow, B. (2012). *God and Necessity*. Oxford: Oxford University Press.

Leite, A. (2010). Fallibilism. In J. Dancy, E. Sosa, & M. Steup (Eds.) *A Companion to Epistemology*, (pp. 370–375). Malden, MA.: Wiley-Blackwell.

Leitgeb, H. (2007). Beliefs in conditionals vs. conditional beliefs. *Topoi*, *26*(1), 115–132.

Leon, F. (2017). From Modal Skepticism to Modal Empiricism. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 247–261). Cham: Springer.

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of tomm. *Cognition*, *50*(1-3), 211–238.

Levi, I. (1988). Iteration of conditionals and the Ramsey test. *Synthese*, *76*(1), 49–81.

——— (1996). *For the sake of the argument: Ramsey test conditionals, inductive inference and nonmonotonic reasoning*. Cambridge: Cambridge University Press.

Levin, J. (2019). A Case for the Method of Cases: Comments on Edouard Machery, *Philosophy Within its Proper Bounds*. *Philosophy and Phenomenological Research*, *98*(1), 230–238.

Lewis, D. K. (1973a). Causation. *The journal of philosophy*, *70*(17), 556–567.

——— (1973b). *Counterfactuals*. Cambridge, MA.: Harvard University Press.

——— (1976). Probabilities of Conditionals and Conditional Probabilities. *The Philosophical Review*, *85*(3), 297–315.

——— (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, *61*(4), 343–377.

——— (1986). *On the Plurality of Worlds*. Oxford: Blackwell Publishers.

Liao, S.-y., & Gendler, T. S. (2019). Imagination. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2019 ed.

Lightner, T. (1997). Hume on conceivability and inconceivability. *Hume Studies*, *23*(1), 113–132.

Linnebo, Ø. (2009). Introduction. *Synthese*, *170*(3), 321–329.

Lowe, E. J. (2012). What is the Source of Our Knowledge of Modal Truths? *Mind*, *121*(484), 919–950.

Lycan, W. G. (2019). *On Evidence in Philosophy*. Oxford: Oxford University Press.

Lyons, J. (2017). Epistemological problems of perception. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2017 ed.

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.

Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Cova, F., Dranseika, V., Ángeles Eraña Lagos, Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., Hristova, E., Jraissati, Y., Kadreva, V., Karasawa, K., Kim, H., Kim, Y., Lee, M., Mauro, C., Mizumoto, M., Moruzzi, S., Olivola, C. Y., Ornelas, J., Osimani, B., Romero, C., Lopez, A. R., Sangoi, M., Sereni, A., Songhorian, S., Sousa, P., Struchiner, N., Tripodi, V., Usui, N., del Mercado, A. V., Volpe, G., Vosgerichian, H. A., Zhang, X., & Zhu, J. (2017). The Gettier Intuition from South America to Asia. *Journal of Indian Council of Philosophical Research*, *34*(3), 517–541.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2018a). Gettier Across Cultures. *Noûs*, *51*(3), 645–664.

Machery, E., Stich, S. P., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2018b). Gettier was framed! In M. Mizumoto, S. P. Stich, & E. McCready (Eds.) *Epistemology for the rest of the world*, (pp. 123–148). Oxford: Oxford University Press.

Mackie, J. L. (1980). *The Cement of the Universe*. Oxford: Oxford University Press.

Mackie, P. (2009). *How Things Might Have Been*. Oxford: Oxford University Press.

Macpherson, F., & Dorsch, F. (Eds.) (2018). *Perceptual Imagination and Perceptual Memory*. Oxford: Oxford University Press.

Maier, J. (2015). The agentive modalities. *Philosophy and Phenomenological Research*, *90*(1), 113–134.

Mallozzi, A. (2018a). Putting modal metaphysics first. *Synthese*, (forthcoming), 1–20. https://doi.org/10.1007/s11229-018-1828-2.

——— (2018b). Two notions of metaphysical modality. *Synthese*, (forthcoming), 1–22. https://doi.org/10.1007/s11229-018-1702-2.

——— (2019). Special issue of synthese on new directions in the epistemology of modality: introduction. *Synthese*, (forthcoming), 1–19. https://doi.org/10.1007/s11229-019-02358-8.

Malmgren, A.-S. (2011). Rationalism and the Content of Intuitive Judgements. *Mind*, *120*(478), 263–327.

Markman, E. M. (1989). *Categorization and Naming in Children*. Cambridge, MA.: MIT Press.

McLeod, S. (2005). Recent Work: Modal Epistemology. *Philosophical Books*, *46*, 235–245.

Medin, D. L. (1989). Concepts and conceptual structure. *American psychologist*, *44*(12), 1469–1481.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.) *Similarity and Analogical Reasoning*, (pp. 179–195). Cambridge: Cambridge University Press.

Menzies, P., & Beebee, H. (2019). Counterfactual theories of causation. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2019 ed.

Mill, J. S. (1882). *A system of logic*, vol. 1. New York, NY.: Harper & Brothers, 8th ed.

Millikan, R. G. (2000). *On clear and confused ideas: An essay about substance concepts*. Cambridge: Cambridge University Press.

Mitchell, D. B., & Richman, C. L. (1980). Confirmed Reservations: Mental Travel. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 58–66.

Moore, G. E. (1939). Proof of an External World. *Proceedings of the British Academy*, *25*, 273–300.

Moretti, L., & Piazza, T. (2018). Transmission of Justification and Warrant. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2018 ed.

Morreau, M. (2010). It Simply Doesn't Add Up: Trouble with Overall Similarity. *The Journal of Philosophy*, *107*, 469–490.

Moser, P. K. (Ed.) (2002). *The Oxford Handbook of Epistemology*. Oxford: Oxford University Press.

Mumford, S. (2005). Kinds, essences, powers. *Ratio*, *18*(4), 420–436.

——— (2009). Causal Powers and Capacities. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.) *The Oxford Handbook of Causation*, (pp. 265–278). Oxford: Oxford University Press.

Mumford, S., & Anjum, R. L. (2011). *Getting Causes from Powers*. Oxford: Oxford University Press.

——— (2013). *Causation. A Very Short Introduction*. Oxford: Oxford University Press.

Nagel, J. (2012). Intuitions and experiments: a defense of the case method in epistemology. *Philosophy and Phenomenological Research*, *85*(3), 495–527.

Nagel, J., San Juan, V., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, *129*(3), 652–661.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450.

Nanay, B. (2010). Perception and imagination: amodal perception as mental imagery. *Philosophical Studies*, *150*(2), 239–254.

——— (2011a). Do we see apples as edible? *Pacific Philosophical Quarterly*, *92*(3), 305–322.

——— (2011b). Do We Sense Modalities With Our Sense Modalities? *Ratio*, *24*, 299–310.

Nelson, K. (1974). Concept, Word, and Sentence: Interrelations in Acquisition and Development. *Psychological Review*, *81*(4), 267–285.

Nichols, S. (2006a). Imaginative Blocks and Impossibility: An Essay in Modal Psychology. In S. Nichols (Ed.) *The Architecture of the Imagination*, (pp. 237–255). Oxford: Oxford University Press.

Nichols, S. (Ed.) (2006b). *The Architecture of the Imagination. New Essays on Pretence, Possibility, and Fiction*. Oxford: Oxford University Press.

Nichols, S., & Stich, S. P. (2003). *Mindreading: an integrated account of pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.

Noë, A. (2002). Is the visual world a grand illusion? *Journal of Consciousness Studies*, *9*(5-6), 1–12.

——— (2004). *Action in Perception*. Cambridge, MA.: MIT Press.

Nolan, D. (1997). Impossible Worlds: A Modest Approach. *Notre Dame Journal of Formal Logic*, *38*(4), 535–572.

——— (2011). The Extent of Metaphysical Necessity. *Philosophical Perspectives*, *25*, 311–339.

——— (2017). Naturalised Modal Epistemology. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 7–27). Cham: Springer.

NOS (2017a). De vijf meest gestelde vragen over fipronil, en de antwoorden. Online news article. https://nos.nl/artikel/2187250-de-vijf-meest-gestelde-vragen-over-fipronil-en-de-antwoorden.html (accessed July 30th 2018).

——— (2017b). Voedsel- en Warenautoriteit waarschuwt voor besmette eieren. Online new article. https://nos.nl/artikel/2185885-voedsel-en-warenautoriteit-waarschuwt-voor-besmette-eieren.html (accessed July 30th 2018).

Oden, C., Gregg (1977). Fuzziness in semantic memory: Choosing exemplars of subjective categories. *Memory & Cognition*, *5*(2), 198–204.

Olkhovikov, G. K., & Wansing, H. (2018). An axiomatic system and a tableau calculus for stit imagination logic. *Journal of Philosophical Logic*, *47*(2), 259–279.

——— (2019). Simplified tableaux for stit imagination logic. *Journal of Philosophical Logic*, *48*, 981–1001.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Orne, M. T. (1962). On the Social Psychology of Psychological Experiment: With Particular Reference to Demand Characteristics and their Implications. *American Psychologist*, *17*(11), 776–783.

Papineau, D. (1992). Reliabilism, Induction and Scepticism. *The Philosophical Quarterly*, *42*(166), 1–20.

Pappas, G. (2017). Internalist vs. externalist conceptions of epistemic justification. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, fall 2017 ed.

Pappas, S., George (Ed.) (1979). *Justification and Knowledge*. Dordrecht: D. Reidel Publishing Company.

Paul, L. (2009). Counterfactual Theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.) *The Oxford Handbook of Causation*, (pp. 158–184). Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York, NY.: Cambridge University Press.

Pezzulo, G. (2011). Grounding Procedural and Declarative Knowledge in Sensorimotor Anticipation. *Mind & Language*, *26*(1), 78–114.

——— (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger, & W. Wiese (Eds.) *Philosophy and Predictive Processing*, (pp. 1–20). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573215.

Pezzulo, G., Barca, L., Bocconi, A. L., & Borghi, A. M. (2010). When affordances climb into your mind: advantages of motor simulation in a memory task performed by novice and expert rock climbers. *Brain and Cognition*, *73*(1), 68–73.

Pezzulo, G., & Castelfranchi, C. (2009). Thinking as the control of imagination: a conceptual framework for goal-directed systems. *Psychological Research*, *73*, 559–577.

Pezzulo, G., & Cisek, P. (2016). Navigating the Affordance Landscape: Feedback Control as a Process Model of Behavior and Cognition. *Trends in Cognitive Science*, *20*(6), 414–424.

Philie, P. (2009). Entitlement as a response to I-II-III scepticism. *Synthese*, *171*, 459–466.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *PNAS*, *114*(18), 4649–4654.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65–94.

Phillips, J., Morris, A., & Cushman, F. (2019). How We Know What Not To Think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040.

Popper, K. (1959). On the Use and Misuse of Imaginary Experiments, especially in Quantum Theory. In *The Logic of Scientific Discovery*, (pp. 442–456). London: Hutchinson.

Prasada, S. (2000). Acquiring generic knowledge. *Trends in cognitive sciences*, *4*(2), 66–72.

Price, H. (1990). Why 'Not'? *Mind*, *99*(394), 221–238.

Priest, G. (1997). Sylvan's Box. *Notre Dame Journal of Formal Logic*, *38*, 573–581.

——— (1998). What is so Bad about Contradictions? *The Journal of Philosophy*, *95*(8), 410–426.

——— (2016). *Towards Non-Being*. Oxford: Oxford University Press, 2nd ed.

——— (2018). Metaphysical necessity: a skeptical perspective. *Synthese*, (forthcoming), 1–13. https://doi.org/10.1007/s11229-018-1885-6.

Priest, G., Berto, F., & Weber, Z. (2018). Dialetheism. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, fall 2018 ed.

Prinz, J. (2004). *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, MA.: MIT Press.

Pryor, J. (2000). The skeptic and the dogmatist. *Noûs*, *34*(4), 517–549.

Psillos, S. (2009). Regularity Theories. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.) *The Oxford Handbook of Causation*, (pp. 131–157). Oxford: Oxford University Press.

Putnam, H. (1973). Meaning and reference. *Journal of Philosophy*, *70*, 699–711.

——— (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, *7*, 131–193.

——— (1990). Is water necessarily H_2O? In J. Conant (Ed.) *Realism with a human face*, (pp. 54–79). Cambridge, MA.: Harvard University Press.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, *25*(2), 157–182.

Quine, W. V. O. (1948). On What There Is. *The Review of Metaphysics*, *2*(5), 21–38.

——— (1960). *Word & Object*. Cambridge, MA.: The MIT Press.

——— (1969). Natural Kinds. In N. Rescher (Ed.) *Essays in honor of Carl G. Hempel*, (pp. 5 –23). Dordrecht: Springer.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010). Counterfactual Reasoning: Developing a Sense of 'Nearest Possible World'. *Child Development*, *81*(1), 376–389.

Ramsey, F. P. (1929). General Propositions and Causality. In R. Braithwaite (Ed.) *The Foundations of Mathematics and Other Logical Essays*, (pp. 237–255). Eastford CT.: Martino Publishing, 2013 ed.

Redshaw, J., Leamy, T., Pincus, P., & Suddendorf, T. (2018). Young children's capacity to imagine and prepare for certain and uncertain future outcomes. *PloS one*, *13*(9), 1–12.

Richman, C. L., Mitchell, D. B., & Reznick, J. S. (1979). Mental Travel: Some Reservations. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(1), 13–18.

Ripley, D. (2012). Structures and circumstances: two ways to fine-grain propositions. *Synthese*, *189*, 97–118.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony (Eds.) *Similarity and analogical reasoning*, (pp. 21–59). New York, NY.: Cambridge University Press.

Robinson, K. S. (2020). The Coronavirus is Rewriting our Imaginations. The New Yorker. https://www.newyorker.com/culture/annals-of-inquiry/the-coronavirus-and-our-future (accessed May 26th 2020).

Roca-Royes, S. (2007). Mind Independence and Modal Empiricism. In C. Penco, M. Vignolo, V. Ottonelli, & C. Amoretti (Eds.) *4th Latin Meeting in Analytic Philosophy*, (pp. 117–135). CEUR Workshop Proceedings.

——— (2010). Modal Epistemology, Modal Concepts and the Integration Challenge. *Dialectica*, *64*(3), 335–361.

——— (2011a). Conceivability and *De Re* Modal Knowledge. *Noûs*, *45*(1), 22–49.

——— (2011b). Modal knowledge and counterfactual knowledge. *Logique et analyse*, *54*(216), 537–552.

——— (2017). Similarity and Possibility: An Epistemology of *de re* Possibility for Concrete Entities. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 221–245). Cham: Springer.

——— (2019a). Concepts and the Epistemology of Essence. *Dialectica*, *73*(1-2), 3–29.

——— (2019b). Rethinking the epistemology of modality for *abstracta*. In I. Fred, & J. Leech (Eds.) *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, (pp. 246–265). Oxford: Oxford University Press.

——— (forthcoming). The integration challenge. In O. Bueno, & S. Shalkowski (Eds.) *The Routledge Handbook of Modality*. New York, NY.: Routledge.

Rosch, E. (1978). Principles of Categorization. In E. Roach, & L. B. B. (Eds.) *Cognition and Categorization*, (pp. 27–48). Hillsdale, NJ.: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rosnow, R. L. (2002). The Nature and Role of Demand Characteristics in Scientific Inquiry. *Prevention & Treatment*, *5*, 1–7.

Russell, B. (1948). *Human Knowledge: Its Scope and Limits*. New York, NY.: Simon and Schuster.

Russell, S. (1988). Analogy by Similarity. In D. H. Helman (Ed.) *Analogical Reasoning*, (pp. 251–269). Dordrecht: Kluwer Academic Publishers.

Rysiew, P. (2020). Naturalism in epistemology. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2020 ed.

Salmon, N. (1981). *Reference and Essence*. Princeton, NJ.: Princeton University Press.

Sayadsayamdost, H. (2015). On Normativity and Epistemic Intuitions: Failure of Replication. *Episteme*, *12*(1), 95–116.

Schaffer, J. (2016). The metaphysics of causation. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, fall 2016 ed.

Schoonen, T. (2020). The Problem of Modally Bad Company. *Res Philosophica*, (forthcoming), 1–22. https://doi.org/10.11612/resphil.2020.97.4.0000.

Searle, J. (1980). Minds, Brain and Programs. *Behavioral and Brain Sciences*, *3*, 417–457.

Shalkowski, S. A. (1992). Supervenience and Causal Necessity. *Synthese*, *90*, 55–87.

Shapiro, L. (2007). The Embodied Cognition Research Programme. *Philosophy Compass*, *2*(2), 338–346.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Shoemaker, S. (1998). Causal and Metaphysical Necessity. *Pacific Philosophical Quarterly*, *79*(1), 59–77.

Shope, R. (1983). *The Analysis of Knowing: A Decade of Research*. Princeton, NJ.: Princeton University Press.

Shtulman, A., & Phillips, J. (2018). Differentiating 'could' from 'should': Developmental changes in modal cognition. *Journal of Experimental Child Psychology*, *165*, 161–182.

Sidelle, A. (1989). *Necessity, Essence, and Individuation: A Defense of Conventionalism*. Ithaca, NY.: Cornell University Press.

Sider, T. (2013). *Writing the Book of the World*. Oxford: Oxford University Press.

Siegel, S. (2006). Which Properties are Represented in Perception? In T. S. Gendler, & J. Hawthorne (Eds.) *Perceptual Experience*, (pp. 481–503). Oxford: Oxford University Press.

——— (2016). The Contents of Perception. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2016 ed.

Sjölin Wirling, Y. (2019a). An integrative design? How liberalised modal empiricism fails the integration challenge. *Synthese*, (forthcoming), 1–19. https://doi.org/10.1007/s11229-019-02426-z.

——— (2019b). *Modal Empiricism Made Difficult*. Ph.D. thesis, University of Göteborg, Göteborg, Sweden.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Cambridge, MA.: Harvard university Press.

Smith, E. E., & Sloman, S. A. (1994). Similarity- versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386.

Sorensen, R. A. (1992a). Thought Experiments And The Epistemology of Laws. *Canadian Journal of Philosophy*, *22*(1), 15–44.

——— (1992b). *Thought Experminents*. Oxford: Oxford University Press.

——— (2002). The Art of the Impossible. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 337–368). Oxford: Oxford University Press.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, *132*, 99–107.

——— (2017). The Metaphysical Gettier Problem and the X-Phi Critique. In R. Borges, C. De Almeida, & P. D. Klein (Eds.) *Explaining Knowledge: New Essays on the Gettier Problem*, (pp. 231–241). Oxford: Oxford University Press.

Spivey, M. J., & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects. *Psychological Research*, *65*, 235–241.

Stalnaker, R. C. (1968). A theory of conditionals. In W. Harper, R. Stalnaker, & G. Pearce (Eds.) *Ifs: conditionals, belief, decision, chance, and time*, (pp. 41–55). Dordrecht: D. Reidel Publishing Company.

Starmans, C., & Friedman, O. (2012). The Folk Conception of Knowledge. *Cognition*, *124*, 272–283.

Stein, E. (1996). *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Oxford: Claredon Press.

Steiner, P. (2014). Enacting Anti-Representationalism: The scope and the limits of enactive critiques of representationalism. *Avant*, *5*(2), 43–86.

Steup, M., & Neta, R. (2020). Epistemology. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, summer 2020 ed.

Steup, M., & Sosa, E. (Eds.) (2005). *Contemporary Debates in Epistemology*. Malden, MA.: Blackwell Publishing.

Stoljar, D. (2007). II—Two Conceivability Arguments Compared. *Proceedings of the Aristotelian Society*, *107*(1), 27–44.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, *74*(2), 149–175.

Strevenson, L. (2003). Twelve conceptions of imagination. *British Journal of Aesthetics*, *43*(3), 238–259.

Strohminger, M. (2015). Perceptual Knowledge of Nonactual Possibilities. *Philosophical Perspectives*, *29*, 363–375.

Strohminger, M., & Yli-Vakkuri, J. (2017). The Epistemology of Modality. *Analysis*, *77*(4), 825–838.

——— (2018a). Knowledge of objective modality. *Philosophical Studies*, *176*, 1155–1175.

——— (2018b). Moderate Modal Skepticism. In M. A. Benton, J. Hawthorne, & D. Rabinowitz (Eds.) *Knowledge, Belief, and God*, (pp. 302–321). Oxford: Oxford University Press.

Stuart, M. T. (2019). Towards a dual process epistemology of imagination. *Synthese*, (forthcoming), 1–22. https://doi.org/10.1007/s11229-019-02116-w.

——— (2020). The Productive Anarchy of Scientific Imagination. *Philosophy of Science*, (forthcoming), 1–22.

Stuart, M. T., Fehige, Y., & Brown, J. R. (Eds.) (2018a). *The Routledge Companion to Thought Experiments*. New York, NY.: Routledge.

Stuart, M. T., Fehige, Y., & Brown, J. R. (2018b). Thought Experiments. State of the Art. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.) *The Routledge Companion to Thought Experiments*, (pp. 1–28). New York, NY.: Routledge.

Sutton, J. (2007). *Without Justification*. Cambridge, MA.: The MIT Press.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and phenomenological research*, *76*(1), 138–155.

Tahko, T. E. (2012). Counterfactuals and Modal Epistemology. *Grazer Philosophische Studien*, *86*, 93–115.

——— (2017). Empirically-Informed Modal Rationalism. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 29–46). Cham: Springer.

——— (2018). The Epistemology of Essence. In A. Carruth, S. Gibb, & J. Heil (Eds.) *Ontology, Modality, Mind: Themes from the Metaphysics of E.J. Lowe*, (pp. 93–110). Oxford: Oxford University Press.

Tennant, N. (2017). Logicism and neologicism. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2017 ed.

Thomas, N. J. T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science*, *23*(2), 207–245.

——— (2018). Mental Imagery. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring ed.

Thomasson, A. (2013). Norms and necessity. *Southern Journal of Philosophy*, *51*(2), 143–160.

Thomson, J. J. (1971). A defense of abortion. *Philosophy and public affairs*, *1*(1), 47–66.

Tidman, P. (1994). Conceivability as a test for possibility. *American Philosophical Quarterly*, *31*(4), 297–309.

Tobin, E. (2010). Crosscutting Natural Kinds and the Hierarchy Thesis. In H. Beebee, & N. Sabbarton-Leary (Eds.) *The Semantics and Metaphysics of Natural Kinds*, (pp. 179–191). New York, NY.: Routledge.

Tooming, U. (2018). There is Something about the Image: A Defence of the Two-Component View of Imagination. *Dialectica*, *72*(1), 121–139.

Turri, J. (2013). A conspicuous art: putting Gettier to the test. *Philosopher's Imprint*, *13*(10), 1–16.

——— (2019). Experimental Epistemology and "Gettier" Cases. In S. Hetherington (Ed.) *The Gettier Problem*, (pp. 199–217). Cambridge: Cambridge University Press.

Turri, J., Blouw, P., & Buckwalter, W. (2015). Knowledge and Luck. *Psychonomic Bulletin and Review*, *22*, 378–390.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131.

——— (1983). Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Vaidya, A. J. (2016). The Epistemology of Modality. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, winter 2016 ed.

——— (2017). Modal Knowledge: Beyond Rationalism and Empiricism. In B. Fischer, & F. Leon (Eds.) *Modal Epistemology After Rationalism*, vol. 378, (pp. 85–114). Cham: Springer.

Vaidya, A. J., & Wallner, M. (2018). The epistemology of modality and the problem of modal epistemic friction. *Synthese*, (forthcoming), 1–27. https://doi.org/10.1007/s11229-018-1860-2.

Van Cleve, J. (1983). Conceivability and the Cartesian argument for Dualism. *Pacific Philosophical Quarterly*, *64*(1), 35–45.

——— (1984). Reliability, Justification, and the Problem of Induction. *Midwest Studies in Philosophy*, *9*(1), 555–567.

Van Inwagen, P. (1998). Modal Epistemology. *Philosophical Studies*, *92*(1/2), 67–84.

Van Leeuwen, N. (2011). Imagination is where the action is. *The Journal of Philosophy*, *108*(2), 55–77.

——— (2013). The Meanings of 'Imagine' Part I: Constructive Imagination. *Philosophy Compass*, *8*(3), 220–230.

Vetter, B. (2015). *Potentiality*. Oxford: Oxford University Press.

——— (2017). Williamsonian modal epistemology, possibility-based. In J. Yli-Vakkuri, & M. McCullagh (Eds.) *Williamson on Modality*, (pp. 314–343). London: Routledge.

Vogel, J. (2000). Reliabilism leveled. *The Journal of Philosophy*, *97*(11), 602–623.

Vosniadou, S., & Ortony, A. (Eds.) (1989a). *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press.

Vosniadou, S., & Ortony, A. (1989b). Similarity and analogical reasoning: a synthesis. In S. Vosniadou, & A. Ortony (Eds.) *Similarity and Analogical Reasoning*, (pp. 1–17). Cambridge: Cambridge University Press.

Vranas, P. B. (2000). Gigerenzer's normative critique of Kahneman and Tversky. *Cognition*, *76*(3), 179–193.

Walker, C. M., & Gopnik, A. (2013). Causality and Imagination. In M. Taylor (Ed.) *The Oxford Handbook of the Development of Imagination*, (pp. 342–358). Oxford: Oxford University Press.

Walton, K. (1990). *Mimesis as Make-Believe*. Cambridge, MA.: Harvard University Press.

Wansing, H. (2017). Remarks on the logic of imagination. A step towards understanding doxastic control through imagination. *Synthese*, *194*(8), 2843–2861.

Weatherson, B. (2004). Morality, fiction, and possibility. *Philosophers' Imprint*, *4*(3), 1–27.

Weinberg, J. M. (2007). How to Challenge Intuitions Empirically Without Risking Skepticism. *Midwest Studies in Philosophy*, *31*, 318–343.

——— (2017). Knowledge, Noise, and Curve-Fitting. In R. Borges, C. de Almeida, & P. D. Klein (Eds.) *Explaining Knowledge. New Essays on the Gettier Problem*, (pp. 253–272). Oxford: Oxford University Press.

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical topics*, *29*(1/2), 429–460.

Wheeler, G. (2020). Bounded Rationality. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2020 ed.

White, A. (1990). *The language of imagination*. Oxford: Blackwell Publishing.

Williamson, T. (2005). The Presidential Address: Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking. *Proceedings of the Aristotelian Society*, *105*, 1–23.

——— (2007). *The Philosophy of Philosophy*. Oxford: Blackwell Publishing.

——— (2014). What is naturalism? In M. C. Haug (Ed.) *Philosophical Methodology: The Armchair or the Laboratory?*, (pp. 29–31). London: Routledge.

——— (2016a). Knowing by Imagining. In A. Kind, & P. Kung (Eds.) *Knowledge Through Imagination*, (pp. 113–123). Oxford: Oxford University Press.

——— (2016b). Modal science. *Canadian Journal of Philosophy*, *46*(4-5), 453–492.

——— (2016c). Philosophical Criticisms of Experimental Philosophy. In J. Sytsma, & W. Buckwalter (Eds.) *A Companion to Experimental Philosophy*, (pp. 22–36). Malden, MA.: Wiley-Blackwell.

——— (2018). Counterpossibles. *Topoi*, *37*, 357–368.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Wilson, R. A., & Foglia, L. (2017). Embodied cognition. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Stanford, CA.: The Metaphysics Research Lab, Stanford University, spring 2017 ed.

Wiltsher, N. (2016). Against the Additive View of Imagination. *Australasian Journal of Philosophy*, *94*(2), 266–282.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul. Translation by C.K. Ogden. Reprinted by Dover Publications, 1999.

——— (1953). *Philosophical Investigations*. Wiley-Blackwell, (2009) revised 4$^{\mathrm{rd}}$ ed. Translated by G.E.M. Anscombe, P.M.S. Hacker, and Joachim Schulte.

——— (1958). *The Blue and Brown Books*. Malden, MA.: Blackwell Publishing.

——— (1967). *Zettel*. Berkeley, CA.: University of California Press. Translated by G.E.M. Anscombe.

Wolpert, D., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

Worley, S. (2003). Conceivability, Possibility and Physicalism. *Analysis*, *63*(1), 15–23.

van Woudenberg, R. (2006). Conceivability and Modal Knowledge. *Metaphilosophy*, *37*(2), 210–221.

Wright, C. (2002). The Conceivability of Naturalism. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 401–439). Oxford: Oxford University Press.

——— (2004). Warrant for nothing (and foundations for free)? In *Aristotelian Society Supplementary Volume*, vol. 78, (pp. 167–212). Wiley Online Library.

——— (2014). On Epistemic Entitlement (II): Welfare Sate Epistemology. In D. D. Zardini (Ed.) *Scepticism and Perceptual Justification*, (pp. 213–247). Oxford: Oxford University Press.

——— (2018). Counter-Conceivability Again. In I. Fred-Rivera, & J. Leech (Eds.) *Being Necessary: Themes of Ontology and Modality from the Work of Bob Hale*, (pp. 266–282). Oxford: Oxford University Press.

Wright, J. C. (2010). On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition*, *115*(3), 491–503.

——— (2013). Tracking instability in our philosophical judgments: Is it intuitive? *Philosophical Psychology*, *26*(4), 485–501.

Yablo, S. (1993). Is Conceivability a Guide to Possibility? *Philosophy and Phenomenological Research*, *53*(1), 1–42.

——— (2002). Coulda, woulda, shoulda. In T. S. Gendler, & J. Hawthorne (Eds.) *Conceivability and Possibility*, (pp. 441–492). Oxford: Oxford University Press.

——— (2014). *Aboutness*. Princeton, NJ.: Princeton University Press.

Yli-Vakkuri, J. (2013). Modal skepticism and counterfactual knowledge. *Philosophical Studies*, *162*, 605–623.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery-congenital aphantasia. *Cortex*, *73*, 378–380.

——— (2016). Reflections on aphantasia. *Cortex*, *74*, 336–337.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heuerman-Williamson, B., Jones, K., MacKisack, M., & Winlove, C. (2020). Phantasia—The psychological significance of lifelong visual imagery vividness extremes. *Cortex*, *130*, 426–440.

Samenvatting

Verhalen van gelijkenis en verbeelding

Een bescheiden epistemologie van mogelijkheden

Dit proefschrift draagt bij aan het debat over hoe we gerechtvaardigde overtuigingen kunnen hebben over *niet-werkelijke mogelijkheden*. Deze studie evalueert benaderingen van de epistemologie van mogelijkheden die op verbeelding en op gelijkenis gebaseerd zijn, ontwikkelt binnen beide benaderingen nieuwe theorieën en onderzoekt de rol van mogelijkheidsoordelen binnen de filosofie. De rode draad door dit werk is *modale bescheidenheid*: ook al kunnen we gewone mogelijkheidsbeweringen terecht geloven (bijv. dat dit kopje kan breken), dit vermogen is beperkt als het gaat om vergezochte mogelijkheden (bijv. dat er een fysiek duplicaat van mij kan zijn dat geen bewustzijn heeft). Het proefschrift bestaat uit drie delen.

Deel I van de dissertatie onderzoekt de op verbeelding gebaseerde epistemologieën van mogelijkheden – d.w.z. de suggestie dat het kunnen *verbeelden* van iets ons rechtvaardigt in de overtuiging dat datgene mogelijk is. Verschillende prominente interpretaties van wat verbeelding is, worden beoordeeld op hun potentieel als basis voor een op verbeelding gebaseerde epistemologie van mogelijkheden. Daaruit blijkt dat de resulterende theorieën niet bijdragen aan een fundamentele verklaring van onze kennis van niet-werkelijke mogelijkheden. Derhalve wordt er een nieuwe theorie van verbeelding voorgesteld: verbeelding als sensomotorische simulatie. Een epistemologie van mogelijkheden gebaseerd op deze theorie van verbeelding komt tegemoet aan de bezwaren die ingebracht zijn tegen de eerdere theorieën en kan kennis van sommige niet-werkelijke mogelijkheden verklaren.

Deel II beoordeelt de suggestie dat we onze overtuigingen over mogelijkheden rechtvaardigen door overtuigingen over de werkelijke wereld te extrapoleren door te *redeneren op basis van gelijkenis*. Allereerst wordt het begrip ‘relevante gelijkenis’ kritisch geëvalueerd en er wordt beargumenteerd dat menige interpretatie van deze notie problematisch is. Daarom wordt er een nieuwe op gelijkenis gebaseerde epistemologie van mogelijkheden voorgesteld die steunt op het concept van (natuurlijke) soort, ons vermogen om objecten te categoriseren als zijnde van een bepaalde soort en ons daaraan gerelateerd uitbreidend redeneringsvermogen. Er wordt betoogd dat de resulterende epistemologie van mogelijkheden cognitief plausibel, modaal bescheiden en in overeenstemming met de empirische bevindingen is.

Deel III bespreekt hoe de filosofie afhankelijk is van mogelijkheidsoordelen. Een belangrijk bezwaar tegen niet-exceptionistische theorieën van gedachte-experimenten wordt uitgebreid en er wordt een oplossing voorgesteld die de vraag oproept of mensen *filosofisch interessante mogelijkheden* met recht kunnen geloven. Tenslotte wordt beweerd, in tegenstelling tot wat in de experimentele filosofie beweerd wordt, dat mensen wel betrouwbaar kunnen oordelen of filosofisch significante mogelijkheden waar zijn. Echter, vergezochte filosofische gevallen blijven buiten ons epistemisch bereik.

Summary

Tales of Similarity and Imagination

A modest epistemology of possibility

This dissertation advances the debate on how we have justified beliefs about *non-actual possibilities.* It evaluates imagination-based and similarity-based approaches to the epistemology of possibility, develops novel accounts of each of them, and examines the role of possibility-judgements in philosophy itself. This is done over the course of three parts. A common theme throughout this work is *modal modesty*: even though we can come to justifiably believe ordinary possibility claims (e.g., that this cup could break), this ability is limited when it comes to more exotic possibilities (e.g., that there could be an unconscious physical duplicate of me).

Part I explores imagination-based epistemologies of possibility – i.e., the suggestion that being able to *imagine* something provides us with justification for believing its possibility. Different prominent interpretations of imagination are evaluated for their potential as a foundation for an imagination-based epistemology of possibility. It is concluded that these theories are unable to (ultimately) explain our knowledge of non-actual possibilities. A new interpretation of imagination, as sensori-motor simulation, is proposed, which does not succumb to the issues raised against the other theories and can provide us with some knowledge of non-actual possibilities.

Part II assesses the suggestion that we justify beliefs about possibilities by extending our beliefs about the actual world through *similarity reasoning.* First, the notion of relevant similarity is critically evaluated and many interpretations of it are argued to be problematic. A new similarity-based epistemology of possibility is proposed, which relies on the notion of (natural) kind, our ability to categorise objects as being of a particular kind, and our ampliative reasoning skills related to this. The resulting epistemology of possibility is argued to be cognitively plausible, in line with empirical findings, and modally modest.

Part III discusses the reliance of philosophy itself on possibility-judgements. It extends an important objection against non-exceptionalist theories of thought experiments. A solution is proposed, raising the question of whether humans can come to justifiably believe *philosophically interesting possibilities.* It is argued, *pace* experimental philosophers, that we do in fact can reliably judge whether philosophically significant possibilities are true. However, exotic philosophical cases remain out of our epistemic reach.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.

ILLC DS-2015-02: **Mathias Winther Madsen**
The Kid, the Clerk, and the Gambler - Critical Studies in Statistics and Cognitive Science

ILLC DS-2015-03: **Shengyang Zhong**
Orthogonality and Quantum Geometry: Towards a Relational Reconstruction of Quantum Theory

ILLC DS-2015-04: **Sumit Sourabh**
Correspondence and Canonicity in Non-Classical Logic

ILLC DS-2015-05: **Facundo Carreiro**
Fragments of Fixpoint Logics: Automata and Expressiveness

ILLC DS-2016-01: **Ivano A. Ciardelli**
Questions in Logic

ILLC DS-2016-02: **Zoé Christoff**
Dynamic Logics of Networks: Information Flow and the Spread of Opinion

ILLC DS-2016-03: **Fleur Leonie Bouwer**
What do we need to hear a beat? The influence of attention, musical abilities, and accents on the perception of metrical rhythm

ILLC DS-2016-04: **Johannes Marti**
Interpreting Linguistic Behavior with Possible World Models

ILLC DS-2016-05: **Phong Lê**
Learning Vector Representations for Sentences - The Recursive Deep Learning Approach

ILLC DS-2016-06: **Gideon Maillette de Buy Wenniger**
Aligning the Foundations of Hierarchical Statistical Machine Translation

ILLC DS-2016-07: **Andreas van Cranenburgh**
Rich Statistical Parsing and Literary Language

ILLC DS-2016-08: **Florian Speelman**
Position-based Quantum Cryptography and Catalytic Computation

ILLC DS-2016-09: **Teresa Piovesan**
Quantum entanglement: insights via graph parameters and conic optimization

ILLC DS-2016-10: **Paula Henk**
Nonstandard Provability for Peano Arithmetic. A Modal Perspective

ILLC DS-2017-01: **Paolo Galeazzi**
Play Without Regret

ILLC DS-2017-02: **Riccardo Pinosio**
The Logic of Kant's Temporal Continuum

ILLC DS-2017-03: **Matthijs Westera**
Exhaustivity and intonation: a unified theory

ILLC DS-2017-04: **Giovanni Cinà**
Categories for the working modal logician

ILLC DS-2017-05: **Shane Noah Steinert-Threlkeld**
Communication and Computation: New Questions About Compositionality

ILLC DS-2017-06: **Peter Hawke**
The Problem of Epistemic Relevance

ILLC DS-2017-07: **Aybüke Özgün**
Evidence in Epistemic Logic: A Topological Perspective

ILLC DS-2017-08: **Raquel Garrido Alhama**
Computational Modelling of Artificial Language Learning: Retention, Recognition & Recurrence

ILLC DS-2017-09: **Miloš Stanojević**
Permutation Forests for Modeling Word Order in Machine Translation

ILLC DS-2018-01: **Berit Janssen**
Retained or Lost in Transmission? Analyzing and Predicting Stability in Dutch Folk Songs

ILLC DS-2018-02: **Hugo Huurdeman**
Supporting the Complex Dynamics of the Information Seeking Process

ILLC DS-2018-03: **Corina Koolen**
Reading beyond the female: The relationship between perception of author gender and literary quality

ILLC DS-2018-04: **Jelle Bruineberg**
Anticipating Affordances: Intentionality in self-organizing brain-body-environment systems

ILLC DS-2018-05: **Joachim Daiber**
Typologically Robust Statistical Machine Translation: Understanding and Exploiting Differences and Similarities Between Languages in Machine Translation

ILLC DS-2018-06: **Thomas Brochhagen**
Signaling under Uncertainty

ILLC DS-2018-07: **Julian Schlöder**
Assertion and Rejection

ILLC DS-2018-08: **Srinivasan Arunachalam**
Quantum Algorithms and Learning Theory

ILLC DS-2018-09: **Hugo de Holanda Cunha Nobrega**
Games for functions: Baire classes, Weihrauch degrees, transfinite computations, and ranks

ILLC DS-2018-10: **Chenwei Shi**
Reason to Believe

ILLC DS-2018-11: **Malvin Gattinger**
New Directions in Model Checking Dynamic Epistemic Logic

ILLC DS-2018-12: **Julia Ilin**
Filtration Revisited: Lattices of Stable Non-Classical Logics

ILLC DS-2018-13: **Jeroen Zuiddam**
Algebraic complexity, asymptotic spectra and entanglement polytopes

ILLC DS-2019-01: **Carlos Vaquero**
What Makes A Performer Unique? Idiosyncrasies and commonalities in expressive music performance

ILLC DS-2019-02: **Jort Bergfeld**
Quantum logics for expressing and proving the correctness of quantum programs

ILLC DS-2019-03: **András Gilyén**
Quantum Singular Value Transformation & Its Algorithmic Applications

ILLC DS-2019-04: **Lorenzo Galeotti**
The theory of the generalised real numbers and other topics in logic

ILLC DS-2019-05: **Nadine Theiler**
Taking a unified perspective: Resolutions and highlighting in the semantics of attitudes and particles

ILLC DS-2019-06: **Peter T.S. van der Gulik**
Considerations in Evolutionary Biochemistry

ILLC DS-2019-07: **Frederik Möllerström Lauridsen**
Cuts and Completions: Algebraic aspects of structural proof theory

ILLC DS-2020-01: **Mostafa Dehghani**
Learning with Imperfect Supervision for Language Understanding

ILLC DS-2020-02: **Koen Groenland**
Quantum protocols for few-qubit devices

ILLC DS-2020-03: **Jouke Witteveen**
Parameterized Analysis of Complexity

ILLC DS-2020-04: **Joran van Apeldoorn**
A Quantum View on Convex Optimization

ILLC DS-2020-05: **Tom Bannink**
Quantum and stochastic processes

ILLC DS-2020-06: **Dieuwke Hupkes**
Hierarchy and interpretability in neural models of language processing

ILLC DS-2020-07: **Ana Lucia Vargas Sandoval**
On the Path to the Truth: Logical & Computational Aspects of Learning

ILLC DS-2020-08: **Philip Schulz**
Latent Variable Models for Machine Translation and How to Learn Them

ILLC DS-2020-09: **Jasmijn Bastings**
A Tale of Two Sequences: Interpretable and Linguistically-Informed Deep Learning for Natural Language Processing

ILLC DS-2020-10: **Arnold Kochari**
Perceiving and communicating magnitudes: Behavioral and electrophysiological studies

ILLC DS-2020-11: **Marco Del Tredici**
Linguistic Variation in Online Communities: A Computational Perspective

ILLC DS-2020-12: **Bastiaan van der Weij**
Experienced listeners: Modeling the influence of long-term musical exposure on rhythm perception

ILLC DS-2020-13: **Thom van Gessel**
Questions in Context

ILLC DS-2020-14: **Gianluca Grilletti**
Questions & Quantification: A study of first order inquisitive logic