A Matter of Trust

Dynamic Attitudes in Epistemic Logic

Ben Rodenhäuser

A Matter of Trust

Dynamic Attitudes in Epistemic Logic

ILLC Dissertation Series DS-2014-04

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
fax: +31-20-525 5206
illc@uva.nl
www.illc.uva.nl

A Matter of Trust

Dynamic Attitudes in Epistemic Logic

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op donderdag 19 juni 2014, te 10.00 uur

door

Leif Benjamin Rodenhäuser

geboren te Berlijn, Duitsland.

Promotiecommissie

Promotores:
Prof. dr. F.J.M.M. Veltman
Prof. dr. L.C. Verbrugge

Co-promotor:
Dr. S. Smets

Overige leden:
Prof. dr. J.F.A.K. van Benthem
Prof. dr. J.A.G. Groenendijk
Dr. W.H. Holliday
Prof. dr. L.S. Moss
Prof. dr. H. Rott

Faculteit der Natuurwetenschappen, Wiskunde en Informatica
Universiteit van Amsterdam
Science Park 904
1098 XH Amsterdam

Cover design by Julia Ochsenhirt.
Printed and bound by Ipskamp, Enschede.
Cover printed by Spezialdruck, Berlin.

ISBN: 978–94–6259–220–9

Between stimulus and response there is a space. In that space is our power to choose our response. In our response lies our growth and our freedom.

Viktor E. Frankl

Contents

Acknowledgments

The support I have received from my supervisors is gratefully acknowledged: thank you, Sonja Smets, Frank Veltman, and Rineke Verbrugge—for your time, for your patience, for your comments and suggestions. Thank you, especially, Sonja, for working with me over these past four years. Thank you, Johan van Benthem, Jeroen Groenendijk, Wesley Holliday, Larry Moss and Hans Rott for agreeing to serve on my thesis committee. It has been an honour for me to have my work evaluated by you. Thank you to Alexandru Baltag, who inspired me to work in epistemic logic when I was a Master's student at the ILLC, and whose ideas have had the most influence on this dissertation by far. Thank you to the friends and colleagues inside and outside of the ILLC who have influenced this work through discussions and comments: Johan van Benthem, Cédric Dégremont, Jan van Eijck, Virginie Fiutek, Jonathan Ginzburg, Paula Henk, Andreas Herzig, Sujata Ghosh, Barteld Kooi, Nina Gierasimczuk, Davide Grossi, Eric Pacuit, Bryan Renne, Olivier Roy, Jakub Syzmanik, Zoé Christoff. And thank you to Jenny Batson, Karine Gigengack, Tanja Kassenaar, Peter van Ormondt, Yde Venema and Marco Vervoort, for making the ILLC such a nice place to work at.

Mir liegt es am Herzen, meinen besten Freunden zu danken: Dine und Patrick—danke für alles. Ich vermute, euch ist gar nicht bewusst, wie viel ihr in den letzten Jahren für mich getan habt. Danke, liebe Julia, wie schön, dass ich dich nach so langer Zeit unverhofft wiedertreffen durfte. Mein ganz besonderer Dank gilt meiner Familie, bei der ich mich viel zu selten melde und die trotzdem immer für mich da ist: meiner Mutter, Sabine, meinem Vater, Ulf, meiner Schwester, Caroline, und meinen Großeltern, Heinrich und Juliane.

Die Arbeit ist der Erinnerung an meinen Großvater Eugen gewidmet.

Köln, im Mai 2014
Ben Rodenhäuser

Introduction

While propositional attitudes—like knowledge and belief—capture an agent's opinion about a particular *piece* of information, dynamic attitudes, as understood in this dissertation, capture an agent's opinion about a particular *source* of information, more precisely: they represent the agent's assessment of (or opinion about) the reliability (or trustworthiness) of the source. The project of this dissertation is to study the latter notion from a general qualitative vantage point.

Reliability and trustworthiness are dispositional predicates. If a source is considered reliable by an agent, we interpret this as saying that the agent will consistently rely on information received from that source in concrete scenarios; he or she will have a disposition to *believe* the source, and this in turn means—to a first approximation—that the agent will *come to believe* what the source says.

Characterizations like this are inherently dynamic: they are phrased in terms of changes the epistemic state of the agent undergoes upon receipt of information from a particular source.[1] This simple observation paves the way for formally representing reliability assessments using a familiar format, namely, in terms of operators that transform information states given informational inputs. Such operators generalize the notion of a *belief revision policy*, taken to represent a generic way of "coming to believe," a strategy for accepting the content of a particular informational input.[2]

Usually, how reliable a source is assumed to be will depend on contextual features. Take the example of a source that is a mathematician. Such a source may be considered extremely reliable when speaking to mathematical mat-

[1] I will use the terms "epistemic state" and "information state" interchangeably. Notice that, in this dissertation, these are *not* formally defined terms. The correct picture in reading this dissertation is that epistemic states/information states are *captured* by formal structures of a certain kind. Plausibility orders (cf. §1.1) will play a major role here. But notice that, conceptually speaking, the way an agent assesses the reliability of a source is also part of her epistemic "make-up", i.e., her information state.

[2] Cf. van Benthem and Martinez (2008), Baltag and Smets (2008), Baltag, van Ditmarsch, and Moss (2008) for more discussion of the notion of a belief revision policy.

ters, but unreliable when speaking to matters of organizing daily life. So we may trust a particular source on particular matters, but not in general, on all matters.[3] But clearly, the idea of representing reliability assessments in terms of changes of information states is general enough to handle both cases, the simplistic case of a "uniform" reliability assessment, and the more realistic case of a "mixed", contextually dependent one.

As a more concrete entry point to our topic, consider an example from Spohn (2009), who asks us to consider various ways in which I could receive the information that there are tigers in the Amazon jungle:

— *I read a somewhat sensationalist coverage in the yellow press claiming this.*
— *I read a serious article in a serious newspaper claiming this.*
— *I read the Brazilian government officially announcing that tigers have been discovered in the Amazon area.*
— *I see a documentary on TV claiming to show tigers in the Amazon jungle.*
— *I read an article in Nature by a famous zoologist reporting of tigers there.*
— *I travel to the Amazon jungle, and see the tigers.*

Spohn uses this example to illustrate the idea that "evidence comes more or less firmly": "in all six cases," he writes, "I receive the information that there are tigers in the Amazon jungle, but with varying and, I find, increasing certainty."

One way to gauge the relative firmness in each case is to observe that receiving the information from sources further down in the above list makes receiving the information from sources higher up in the list *redundant*, while the converse does not hold. Having seen the tigers with my own eyes, the sensationalist coverage in the yellow press tells me nothing new. On the other hand, after reading the sensationalist coverage, it will still be informative to actually *see* the tigers. It may prompt me to say: "Gosh, there *really are* tigers in the Amazon jungle."

Another way to make the same case is by noticing that my degree of confidence that there are actually tigers in the Amazon jungle is likely to be higher after reading about them in Nature, than after reading about them in the yellow press. Suppose a travel agent offers me a trip to a safari through the

[3] This point is frequently made in the literature, cf., e.g., Liau (2003), who develops a notion of "trust on a sentence φ" (rather than trust *simpliciter*) and Sperber, Clément, Heintz, Mascaro, Mercier, Origgi, and Wilson (2010), who emphasize that trust is bestowed upon sources "on a particular topic in particular circumstances."

Amazon jungle. Having read about the tigers in Nature, I will be confident that the trip will provide me with an opportunity to see some tigers. But having read about the tigers in the yellow press, chances are I won't be so sure.

Why is that? According to the picture outlined so far, it is the fact that the different sources of information—the yellow press, the Brazilian government, my own eyes, etc.—are *trusted* to different degrees, which is responsible for the differences in firmness. Here, we understand trust in the epistemic-dynamic sense alluded to above: a particular level or degree of trust is given by a particular assessment of reliability; a reliability assessment is encoded by a collection of potential belief changes that capture the disposition of our agent to "change his mind" in a particular way when receiving information from that particular source.[4]

On this view, the evidential firmness is thus not inherent in the content of the evidence received, but derives from the fact that the evidence is received from a particular source, towards which the recipients has a particular (dynamic) attitude. In this sense, as the title of this dissertation suggests, the extent to which information flows, changing our epistemic state, as a source presents us with information, is *a matter of trust*.

The Main Idea. As is already implicit in the preceding remarks, the main idea pursued in this dissertation is to formally develop the notion of a dynamic attitude, as a representation of an assessment of reliability, in the context of, and drawing upon, existing work on *information change*, which has studied the transformation of information states due to the receipt of informational inputs; more concretely, this line of research (or rather: these lines of research, cf. fn. 5) is (are) concerned with the way agents *change their minds* when given new information, or, put differently, how agents incorporate new information into their information state.[5] In adopting this perspective on in-

[4]Our everyday conception of trust is much richer than this. For example, one may trust a business partner not to breech the confidentiality of an agreement, or one may trust a friend to help in the case of an emergency. In both cases, it is trust in another's actions that is at stake (trust that a certain action will not be taken in the case of the business partner, trust that a certain action will be taken in the case of the friend), not trust in the quality of information received from him or her. Trust in another's actions is studied in the frame of the more general "cognitive theory of trust" (Castelfranchi and Falcone 1998). In this dissertation, we will focus exclusively on trust in the epistemic-informational sense outlined in the main text, studying, in Liau's phrasing, "the influence of trust on the assimilation of acquired information into an agent's belief state" (Liau 2003).

[5]The study of information change has a rich tradition. Particularly relevant for this dissertation is the body of work established in belief revision theory (starting with Alchourrón,

formation, information is treated as "something that may enter some belief state and change or transform it into another belief state" (Rott 2008). Thus, we are chiefly interested in what information *does* rather than what information *is*. We do not offer a definition of information, but rather individuate pieces of information in terms of the potential effects of receiving information.[6]

Making use of this picture of information change, we may capture reliability assessments using the following format.[7] Upon receipt of an informational input P, an agent will change her epistemic state in a particular way, moving from some "input state" $\mathcal{S}$ to an "output state" $\mathcal{S}'$ that incorporates the input, where just how the input is incorporated will depend on the extent to which the agent trusts the source. A reliability assessment towards a source may thus be represented by means of a plan to react in a specific way when information from that source is received. That is, a reliability assessment will be given by a function that assigns to each given information state and propositional input a new information state. We shall call such functions *dynamic attitudes*. They can be understood as *strategies for belief change*, encoding how the agent will react when receiving information from a certain source in a certain state. A useful mental representation for such a strategy is as a ternary graph, with the nodes given by information states, and the labels given by informational inputs. If, e.g., the triple $(\mathcal{S}, P, \mathcal{S}')$ is an element of the graph representing an assessment of the trustworthiness of a given source, then we read this as saying that the agent plans to make a transition to the information state $\mathcal{S}'$ upon receipt of the informational input that P in the information state $\mathcal{S}$.

Usually, there will be too many information states and informational inputs to make it economical, or even feasible, to actually draw such a graph. But assuming a sparse ontology in which there are just three information states, $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$, and two possible informational inputs, P and Q, a con-

Gärdenfors, and Makinson (1985), Gärdenfors (1988), cf. also the references in fn. 12.), and dynamic epistemic logic (Plaza 1989, Gerbrandy 1999, Baltag and Moss 2004, van Benthem 2007, Baltag and Smets 2008, van Benthem 2011). The work carried out here has also been influenced by work in dynamic semantics, in particular Veltman (1996), and by the early presentation of the program of "logical dynamics" in van Benthem (1996).

[6] This is somewhat similar to the way in which Stalnaker (1978), in his influential study of assertion, aimed at analyzing the effects of assertions, rather than at providing a definition of what an assertion is.

[7] The following remarks gloss over all the details that matter from a formal perspective: how are "information states" exactly represented mathematically? What exactly is the formal correlate of an "informational input"? etc. Precise definitions for our specific purposes follow in Chapter 1.

crete example of a dynamic attitude—represented in the way just suggested—may look like the diagram in Figure 1.

Of course, to know whether the diagram depicts a "trusting" attitude, we need to know more about the propositional attitudes the agent holds in each of the possible information states. Suppose, for example, that in $\mathcal{S}_3$, the agent believes neither P nor Q, while in $\mathcal{S}_2$ she believes P but not Q, and in $\mathcal{S}_1$ she believes both P and Q. Under this assumption, it is easy to check that upon receiving the information that X in one of the three information states in $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$, the agent comes to believe that X (in the resulting state), where X is one of P and Q. It is then natural to think of the depicted dynamic attitude as a form of trust.

But interestingly, we can also reverse the perspective. Suppose we already know that the agent trusts the source the attitude towards which we are depicting in Figure 1, based on our (pre-formal) knowledge about the modeling domain. If that is so, then it is natural to think that the agent would come to believe what the source tells him. Moreover, it is natural to think that the agent would not change his state if he already believed what the source told him. So we could decide to say that our agent believes that X (with $X \in \{P, Q\}$) iff the information state of the agent (one of $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$) is not changed by receiving the information that X. In other words, we could define a propositional attitude ("belief") as the fixed point of our given dynamic one: belief in X is satisfied in just those information states in which receiving the information that X from a trusted source is redundant. It then turns out that, according to this "definition", the agent believes that P in the information state $\mathcal{S}_1$, and that the agent believes that Q in the information states $\mathcal{S}_1$ and $\mathcal{S}_2$ (which is just what we had stipulated earlier). It also turns out that, indeed, our agent comes to believe that X whenever she receives the information that X from the source (check, for instance, what happens if the agent receives the information that P in state $\mathcal{S}_3$). Notice that not all possible diagrams in the style of the one drawn in Figure 1 satisfy this latter property. What is required for our notion of belief derived from the given dynamic attitude to be reasonable is that the dynamic attitude is *idempotent*: that is, receiving the same informational input from the source twice should have the same effect as receiving it once. Idempotence will play a crucial role in the framework of this dissertation.

Overview. The preceding example illustrates the interplay between dynamic and propositional attitudes that will be a main focus of this dissertation. In fact, the idea of studying this interplay provides a unifying thematic thread for much of the work presented here. Here is an outline of the following

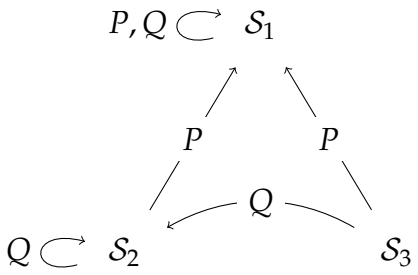


FIGURE 1. Graph representation of a dynamic attitude, picturing an agent's disposition to change her information state when receiving new information, under the assumption that the space of possible information states is given by $\{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$, and the space of possible informational inputs is given by $\{P, Q\}$. For example, the agent plans, if in state $\mathcal{S}_2$, to transform, on receipt of the information that P, her state $\mathcal{S}_2$ into $\mathcal{S}_1$; on the other hand, she plans to remain in state $\mathcal{S}_2$ if she receives the information that Q in state $\mathcal{S}_2$.

chapters that highlights how each chapter is related to this red thread.[8]

In Chapter 1, which introduces the key notions to be studied in the dissertation, we show that (introspective) propositional attitudes naturally arise as *fixed points* of dynamic attitudes; conversely, dynamic attitudes can be seen as chosen with a specific propositional attitude in mind which constitutes the *target of belief change*. We show that the class of dynamic attitudes characterizes the class of introspective propositional attitudes using the fixed point operation; and that a subsumption order on dynamic attitudes can be defined that exactly matches the usual entailment order on propositional attitudes. More generally, our vantage point establishes a systematic link between the static (i.e., propositional) and the dynamic level that we exploit in the remainder of the dissertation.

Chapter 2 studies various forms of trust and distrust, and intermediate forms of "semi-trust". We start out from the notion of acceptance, understood in a broadly Stalnakerian sense: "to accept a proposition is to treat it as a true proposition in one way or another" (Stalnaker 1984).[9] Our discussion

[8] A preliminary version of Chapter and 1 and Chapter 2 appeared in Baltag, Rodenhäuser, and Smets (2012). A preliminary version of Chapter 3 appeared in Rodenhäuser (2013). Chapters 4–6 are new; Chapter 5 is based on joint work with Alexandru Baltag and Sonja Smets.

[9] According to Stalnaker, this characterization picks out a class of propositional attitudes that "are sufficiently like belief in some respects to justify treating them together with it" (ibid.).

complements this static conception of acceptance with a dynamic one, i.e., we study, using our notion of a dynamic attitude, various ways in which an agent may come to treat a certain proposition as true. This approach leads us to identify a number of classes of dynamic attitudes that can be seen as capturing natural ways of assessing the reliability of a source. We identify typical representatives of each class and relate them systematically to the class of propositional attitudes using the notion of a fixed point. We also broaden the focus by considering a multi-agent version of our setting (§§2.7–2.8), which allows us to study properties of communication acts.

Chapter 3 takes on the topic of minimal change that has traditionally played a foundational role in belief revision theory. Compared to the tradition, we widen the focus by not only considering minimal changes to induce belief in a given proposition, but rather studying a general notion of optimality relative to *any* given fixed point. This allows us to further study the question in which sense the typical dynamic attitudes identified in the previous chapter are really special (and, indeed, in many cases, *canonical*, that is, uniquely optimal for their fixed point).

In Chapter 4, we switch the perspective, and study the *preservation* of propositional attitudes under certain classes of transformations. We devote particular attention to preservation under substructures, a form of preservation that has traditionally been important in model theory. From the point of view of information dynamics, the fact that a propositional attitude is preserved under substructures means that the agent's specific opinion (as captured by the propositional attitude) is stable under receipt of hard information. Again, our perspective emphasizes the connection between dynamic and propositional attitudes, as we characterize preservation properties of propositional attitudes in terms of fixed points of dynamic attitudes.

Chapter 5 discusses the link between the static and the dynamic level that has received most attention in the dynamic epistemic logic literature. In this chapter, we study modal languages extended with dynamic modalities and show how the static base language can already define the dynamic part. This allows us to prove generic completeness theorems for our logics. We extend the analysis to a multi-agent version of our setting (drawing on the work of §§2.7–2.8), in which agents figure not only as recipients, but are also sources of information themselves.

The final chapter, Chapter 6, shifts the perspective, providing another perspective on the formal setting developed here: we observe that the dynamic attitudes we have worked with so far can be interpreted not only as reliability assessments on behalf of an agent, but also as denotations for epistemic modals in natural language. Our main point is that the results of this disserta-

tion are also of potential interest to the community working on the semantics of natural language.

Previous Research. Traces of our idea of representing reliability assessments in terms of transformations of information states are found scattered across the literature, sometimes formulated in terms of notions like "evidential reliability" or "epistemic trust".[10] In Bayesian Epistemology, the reliability of a source is captured by weights or probabilities attached to the new information, which determine different ways of processing it (for example, Bayesian conditioning versus Jeffrey conditioning).[11] In Belief Revision theory, various methods for (iterated) belief revision have been proposed that can be understood as corresponding to different attitudes to the incoming information.[12] These operations have also made their way into the dynamic epistemic logic literature, where they have been studied using logical tools.[13]

Another line of resarch relevant for this dissertation stems from the multi-agent systems tradition in Artificial Intelligence, which has led to a body of work on the notion of trust.[14] The work from this tradition that is most closely related to our setting is the paper by Liau (2003), who introduces a modal language extended with trust operators of the form $T_{ab}\varphi$, read as "agent a trusts agent b on φ." Liau then formalizes the idea that if an agent trusts another on a certain piece of information (represented by a sentence of the language), then on receiving that piece of information, the agent comes to believe that φ, an idea which is, in its spirit, closely related to ideas presented in this dissertation (cf. in particular, Chapter 2). While Liau implements the idea in a static setting, this dissertation offers, as outlined above, a dynamic formalization in terms of operations on information states, with the advantage of connecting

[10] E.g., Spohn (2009) studies a variety of revision operations, parametrized by their "evidential force", meant to capture the idea that information one accepts comes in various degrees of "firmness", motivated by examples like the one discussed at the beginning of this introduction. Lehrer and Wagner (1981) suggest to model the trust an agent places in another agent's claims using a notion of "epistemic weight".

[11] Jeffrey (2004), Halpern (2003).

[12] Boutilier (1996), Spohn (1988, 2009), Nayak (1994), Rott (2004, 2009), among others.

[13] One of the first approaches to connecting dynamic epistemic logic with belief revision theory was Aucher (2003). Later, the work of van Benthem (2007) and Baltag and Smets (2008) defined the current standard in the field. All of these papers build on the earlier work of Plaza (1989), Gerbrandy (1999) and Baltag, Moss, and Solecki (1999). As a field of research, dynamic epistemic logic is much broader than just providing "logics for belief revision", but aims at a broader theory of informational dynamics, cf. the recent monograph van Benthem (2011) with many pointers to the literature.

[14] Cf. Castelfranchi and Falcone (1998), Demolombe (2001), Liau (2003), Dastani, Herzig, Hulstijn, and van der Torre (2004), Herzig, Lorini, Hübner, and Vercouter (2010).

directly to existing work in belief revision theory, dynamic epistemic logic, and dynamic semantics.[15]

Approach. Compared to previous work, the approach of this dissertation is characterized by the following distinctive features.

First, we propose a general qualitative-relational (rather than quantitative) setting that allows us to model a much wider class of reliability assessments than just those appropriate for trusted sources. The approach is thus not limited to dynamic attitudes that induce or contract belief, but provides a general format for capturing reliability assessments of various kinds, including forms of distrust, intermediate types of "semi-trust", and "mixtures" of these. As we shall see (cf. in particular Chapter 2), this allows to model a broad variety of phenomena, thereby widening the scope of existing approaches in belief revision theory.

Second, in contrast to existing research in the multi-agent systems tradition (including Liau's paper cited above), we do not restrict attention to a single notion of trust, but rather formalize a large *family* of trusting and trust-like dynamic attitudes whose qualitative strength can be compared using a notion of dynamic entailment/subsumption. In this way, we can make sense of statements according to which one source is trusted *more* than another—an advantage of the present setting shared with Spohn's approach, which, however, has a quantitative rather than qualitative flavour (Spohn 1988).

Third, rather than considering specific belief revision policies in isolation, we *characterize* these policies within our more general setting, clarifying formally in what sense they are special (cf. Chapter 2), and, indeed, unique (cf. Chapter 3). In this sense, our work can be seen as justifying why these policies have received as much attention in the recent literature as they did.[16]

Finally, our approach is characterized by a specific mix of formal tools that is inspired by existing research in belief revision theory, dynamic semantics and dynamic epistemic logic. The structures we work with—*plausibility orders*, i.e., total preorders on some given set of possible worlds—, and the main examples of transformations on these structures, come from belief revision theory.[17] The dynamic semantics tradition has taught us the importance

[15] We address the latter connection, to dynamic semantics, mainly in the final chapter of the dissertation.

[16] The revision policies that come to mind most readily in this regard are Boutilier (1993)'s "minimal revision" and Nayak (1994)'s "lexicographic revision", whose importance for the dynamic epistemic logic literature mainly stems from the influence of van Benthem (2007) and Baltag and Smets (2008).

[17] Plausibility orders have also been important in philosophical logic more generally, due

of studying the *fixed points* of dynamic transformations that will play an important role throughout this dissertation, and the *dynamic entailment (or subsumption) relations* that can be defined in terms of fixed points. From dynamic epistemic logic, finally, we inherit the general focus on the *interplay between the static and the dynamic perspective* explained above: both belief revision theory and dynamic semantics have tended to emphasize transformations of information states at the expense of the structure inherent in the information states themselves; in dynamic epistemic logic, as in this dissertation, both levels are traditionally treated on a par, and figuring out how they are precisely related is a main concern.

to the influence of David Lewis's "sphere semantics" for counterfactuals: systems of spheres are notational variants of plausibility orders (cf. §1.1.4).

Chapter 1.
Dynamic Attitudes

This chapter introduces the setting that we will work with throughout this dissertation. *Plausibility orders* (§1.1) are representations of the epistemic state of an agent; *propositional attitudes* (§1.2) capture static properties of such orders; *upgrades* (§1.3) are relational transformers, corresponding to types of changes that may be applied to plausibility orders. *Uptake operators* (§1.4) are families of upgrades indexed by propositions. Finally, the class of *dynamic attitudes* (§1.5) is given by a subclass of the class of all uptake operators, subject to a number of additional requirements.

Our setting is purely semantic and language-independent, an important fact that we emphasize in §1.6, focusing on a key constraint imposed on dynamic attitudes: idempotence.

Having introduced our formal machinery, we begin the investigation of our setting in §1.7 and §1.8: *fixed points* of dynamic attitudes (§1.7) represent the propositional atttitudes which are realized by applying particular types of transformations. The notion of *subsumption* (§1.8) provides a natural way to compare the strength of given dynamic attitudes, which ties in naturally with the familiar entailment order on propositional attitudes.

1.1. Plausibility Orders

A *set of possible worlds* is a non-empty set. Given a set of possible worlds W, a *proposition* is a subset of W. Intuitively speaking, a proposition P is "a representation of the world as being a certain way" (Stalnaker 1978), which is to say that P is satisfied in the worlds that are "that way" (i.e., as represented by P), and not satisfied in the worlds that are not "that way". Here, the worlds that are that way are simply the members of the set P, and the worlds that are not that way are the worlds that are not members of P. So, formally, a world $w \in W$ *satisfies* P iff $w \in P$, and a world $v \in W$ *does not satisfy* P iff $v \notin P$.

We use the usual notation for set-theoretic operations on propositions, in particular:

$$\neg P := \{w \in W \mid w \notin P\},\ P \cap Q := \{w \in W \mid w \in P \text{ and } w \in Q\}$$

$$P \cup Q := \{w \in W \mid w \in P \text{ or } w \in Q\},\ P \Rightarrow Q := \{w \in W \mid \text{ if } w \in P, \text{ then } w \in Q\}.$$

In the remainder of the dissertation, unless specifically noted otherwise, we assume a fixed, but arbitrary set of possible worlds W as given.

1.1.1. Plausibility Orders. A *plausibility order $\mathcal{S}$ (on W)* is a pair

$$\mathcal{S} := (\mathrm{S}, \leq_{\mathcal{S}}),$$

where $\mathrm{S} \subseteq W$ is a finite set of possible worlds (called the *domain* of $\mathcal{S}$), and $\leq_{\mathcal{S}} \subseteq \mathrm{S} \times \mathrm{S}$ is a total preorder on S, i.e., a transitive and connected (and thus reflexive) relation.[1]

A plausibility order represents the epistemic state of an (implicit) agent. While the set of possible worlds W comprises the totality of all possibilities that are consistent with some basic (unchangeable, context-independent and time-independent) implicit information about the world, over time, the agent will gain more information about the (real state of the) world, information that allows her to cut down the set of possibilities from the initial set W to a proper subset thereof. The latter set, represented by the domain S of a plausibility order, embodies what we call the agent's *hard information*, assumed to be absolutely certain, irrevocable and truthful. Going further, the agent may also possess *soft information*, that is *not* absolutely certain, and subject to revision if the need arises. This information only allows her to hierarchize the possibilities consistent with her hard information according to their subjective "plausibility", but not to actually discard any of them. This relative hierarchy is given by the relation $\leq_{\mathcal{S}}$.[2]

Here, our assumption is that *the smaller the better*, as in, for example, a ranking of soccer teams: the teams who have smaller numbers in the ranking have

[1] A binary relation R on a given set S is *transitive* if for any $w, v, u \in S$: if $(w, v), (v, u) \in R$, then also $(w, u) \in R$; *connected* if for any $w, v \in S$, either $(w, v) \in R$ or $(v, w) \in R$; and *reflexive* if for any $w \in S$: $(w, w) \in R$.

[2] For the distinction between hard and soft information, cf. van Benthem (2007), who illustrates the difference using the example of a card game: the total number of cards in the deck and the cards I hold myself would typically be taken to be hard information. On the other hand, there is also soft information: "I see you smile. This makes it more likely that you hold a trump card, but it does not rule out that you have not got one" (ibid.). My seeing you smile might lead me to believe that you hold a trump card, but this belief is open to revision as further evidence becomes available.

performed better according to the ranking than those with higher numbers (for example, the team in second place is better than the team in third place, etc.); the difference between a plausibility order and your typical ranking of soccer teams is that several worlds may share the same rank. So the fact that $w \leq_{\mathcal{S}} v$ indicates that the world w is *at least as plausible as* the world v (from the perspective of our agent). Taken together, the structure $\mathcal{S}$ represents the epistemic state of an (implicit) agent, at a given point in time and considered in isolation from other agents.

We shall write $w \approx_{\mathcal{S}} v$ to indicate that w and v are *equiplausible*, i.e., $w \approx_{\mathcal{S}} v$ iff both $w \leq_{\mathcal{S}} v$ and $v \leq_{\mathcal{S}} w$.

We also notice that, in case W is finite, there is a special plausibility order $\mathcal{W} = (W, W \times W)$, representing the *initial state of ignorance*, in which the agent has not been able to exclude any world from consideration, and also has not been able to impose any hierarchy on the worlds she considers possible. Another special order is given by $\varnothing = (\varnothing, \varnothing)$: this order represents the *absurd state*, in which the agent has excluded *all* possible worlds from consideration.

Given a plausibility order $\mathcal{S}$, we often use the infix notation for the pre-order $\leq_{\mathcal{S}}$, writing, for example, "$w \leq_{\mathcal{S}} v$" rather than "$(w, v) \in \leq_{\mathcal{S}}$". If $\mathcal{S}$ is clear from the context, we often drop the subscript, writing "$\leq$" rather than "$\leq_{\mathcal{S}}$". Also, we sometimes write "$(w, v) \in \mathcal{S}$" (rather than "$(w, v) \in \leq_{\mathcal{S}}$"); and "$w \in \mathcal{S}$" (rather than "$w \in \mathrm{S}$").

An example of a plausibility order is provided in Figure 2. According to the diagram, the hard information of the agent is currently given by the proposition that the actual world is among the worlds in $\{w_1, \ldots, w_5\}$. We adopt the convention that worlds that are higher on the page are *more plausible* than those that are lower (again, similarly as one would write a ranking of soccer teams: starting with the best teams higher up in the list). So in our example, w_1 is more plausible than w_2, w_2 is more plausible than w_4, etc. On the other hand (unlike in soccer team rankings), two worlds may be equiplausible, and this is represented by drawing them on the same level: so w_2 and w_3, for example, are equiplausible: $w_2 \approx w_3$ (which means, by definition of equiplausibility, that both $w_2 \leq w_3$ and $w_3 \leq w_2$).

1.1.2. Best Worlds. For any proposition P, the *best P-worlds* (or *most plausible P-worlds*) in a plausibility order $\mathcal{S}$, denoted with $\mathrm{best}_{\mathcal{S}} P$, are the $\leq_{\mathcal{S}}$-minimal elements of P, given by the proposition

$$\mathrm{best}_{\mathcal{S}} P := \{w \in P \cap \mathrm{S} \mid \forall v \in P \cap \mathrm{S} : w \leq_{\mathcal{S}} v\}.$$

So the best P-worlds in $\mathcal{S}$ are those P-worlds in S that are at least as good as any P-world in S according to the hierarchy given by $\leq_{\mathcal{S}}$.

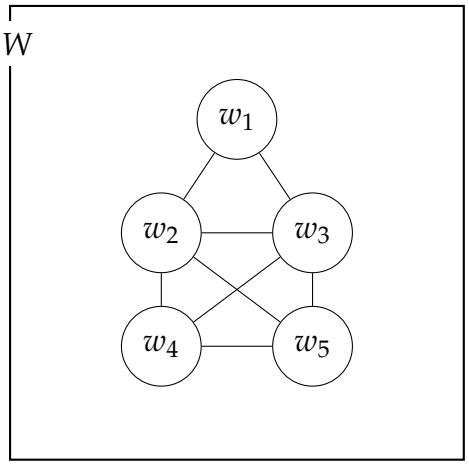


FIGURE 2. A plausibility order with domain $\{w_1, \ldots, w_5\}$. By our drawing convention, worlds that are higher up on the page are more plausible, and worlds on the same level are equiplausible. So, for example, the world w_1 is more plausible than all other worlds.

The *best worlds* (or *most plausible worlds*) in $\mathcal{S}$ are given by $\mathrm{best}\,\mathcal{S} := \mathrm{best}_{\mathcal{S}}\,\mathrm{S}$. Note that $\mathrm{best}\,\mathcal{S} = \varnothing$ iff $\mathrm{S} = \varnothing$. In the example given by Figure 2: $\mathrm{best}\,\mathcal{S} = \{w_1\}$.

1.1.3. UNION, INTERSECTION, CONDITIONALIZATION. For convenience, we lift the set-theoretic operations of intersection and union to plausibility orders. Given plausibility orders $\mathcal{S}_1$ and $\mathcal{S}_2$, the *intersection* $\mathcal{S}_1 \cap \mathcal{S}_2$ of $\mathcal{S}_1$ and $\mathcal{S}_2$, and the *union* $\mathcal{S}_1 \cup \mathcal{S}_2$ of $\mathcal{S}_1$ and $\mathcal{S}_2$ are, respectively, given by the ordered pairs

$$\mathcal{S}_1 \cap \mathcal{S}_2 := (\mathrm{S}_1 \cap \mathrm{S}, \leq_{\mathcal{S}_1} \cap \leq_{\mathcal{S}_2}),\ \mathcal{S}_1 \cup \mathcal{S}_2 := (\mathrm{S}_1 \cup \mathrm{S}, \leq_{\mathcal{S}_1} \cup \leq_{\mathcal{S}_2}).$$

Notice that neither $\mathcal{S}_1 \cup \mathcal{S}_2$ nor $\mathcal{S}_1 \cap \mathcal{S}_2$ is in general guaranteed to be a plausibility order, as we may encounter failures of both transitivity and connectedness. In using these notations in what follows, we will take care to apply them only when given plausibility orders $\mathcal{S}_1, \mathcal{S}_2$ such that $\mathcal{S}_1 \cup \mathcal{S}_2$, $\mathcal{S}_1 \cap \mathcal{S}_2$ *are* plausibility orders.

A more interesting operation on a given plausibility order is what we call "conditionalization". If $\mathcal{S}$ is a plausibility order, and P a proposition, we denote with $\mathcal{S}|_P$ the *conditionalization of* $\mathcal{S}$ *on* P, given by

$$\mathcal{S}|_P := (\mathrm{S} \cap P, \{(w, v) \in \mathcal{S} \mid w, v \in P\}).$$

So conditionalizing a plausibility order $\mathcal{S}$ on a proposition P amounts to simply "throwing away" all the non-P-worlds in S, and restricting the relation

$\leq_S$ accordingly. We may interpret this as corresponding to an event in which the agent receives hard information that P, allowing her to exclude all non-P-worlds from consideration.

1.1.4. Systems of Spheres. A *system of spheres* (sometimes also called an *onion*) is a finite nested set of finite non-empty propositions, that is, a set of sets

$$\mathcal{O} = \{O_1, \ldots, O_n\},$$

where $n \in \omega$, $O_k \subseteq W$ for $1 \leq k \leq n$, and $O_1 \subseteq \ldots \subseteq O_n$. The propositions O_k ($1 \leq k \leq n$) are called *spheres* in $\mathcal{O}$. As with plausibility orders, we allow systems of spheres to be empty.[3] A picture of a system of spheres is provided in Figure 3. In such a drawing, the region covered by each of the circles corresponds to one of the spheres.

For our purposes, systems of spheres will prove useful as equivalent representations of plausibility orders that sometimes allow for a more compact representation in diagrams. The connection between plausibility orders and systems of spheres is as follows. Every plausibility order $\mathcal{S}$ comes equipped with a system of spheres. Given a world $w \in S$, let w^{up} denote

$$w^{\mathrm{up}} := \{x \in S \mid x \leq w\}.$$

The proposition w^{up} collects all worlds that are at least as plausible as w. Now the set of propositions

$$\mathrm{sph}(\mathcal{S}) := \{w^{\mathrm{up}} \mid w \in S\}$$

is a system of spheres, called *the system of spheres for* $\mathcal{S}$. Notice that the innermost sphere of $\mathrm{sph}(\mathcal{S})$ is given by $\mathrm{best}\,\mathcal{S}$.

It is easy to see that the function sph is actually a bijection, so plausibility orders and systems of spheres are in one-to-one correspondence: they are just notational variants. We denote the inverse of sph with ord, that is, for any system of spheres $\mathcal{O}$, $\mathrm{ord}(\mathcal{O})$ is the unique plausibility order $\mathcal{S}$ such that $\mathrm{sph}(\mathcal{S}) = \mathcal{O}$.

Given a system of spheres $\mathcal{O} = \{O_1, \ldots, O_n\}$, we have

$$w \leq_{\mathrm{ord}(\mathcal{O})} v \quad \text{iff} \quad w, v \in O_n \text{ and } \forall O \in \mathcal{O} : v \in O \Rightarrow w \in O.$$

The one-to-one correspondence between plausibility orders and systems of spheres allows us to *define* a plausibility order $\mathcal{S}$ by fixing a system of spheres $\mathcal{O}$; when doing this, it is always understood that $\mathcal{S} := \mathrm{ord}(\mathcal{O})$. To simplify notation, we sometimes write $P \in \mathcal{S}$ to indicate that $P \in \mathrm{sph}(\mathcal{S})$.

[3] Systems of spheres were introduced by Lewis (1973) in the context of his work on counterfactuals. Lewis also noted that a system of spheres in his sense is equivalent to an ordering on possible worlds. The significance of this kind of structure for belief revision theory was realized by Grove (1988).

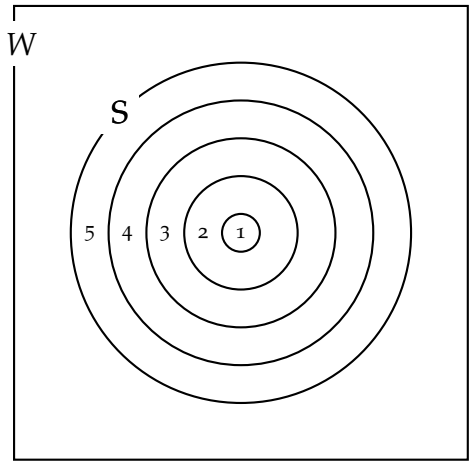


FIGURE 3. A pictorial representation of a system of spheres. The region covered by each circle corresponds to a sphere. The circle labeled 1, for example, gives the *innermost* sphere, and the circle labeled 5 gives the *outermost* sphere. The closer a world is situated towards the center, the more plausible it is in the plausibility order associated with the system of spheres, with the innermost sphere corresponding to the most plausible worlds.

1.1.5. SPOHN ORDINALS. As suggested by our notation, one can assign a natural number to each world w in a given system of spheres $\mathcal{O}$, called the "Spohn ordinal" of w, essentially given by the "rank" of the smallest sphere containing w.

Formally, given a system of spheres $\mathcal{O} = \{O_1, \ldots, O_n\}$, the *Spohn ordinal* $\kappa_{\mathcal{O}}(w)$ of a world $w \in O_n$ is the least natural number k such that $w \in O_k$, i.e.,

$$\kappa_{\mathcal{O}}(w) := \min\{k \in \omega \mid w \in O_k\}.$$

It also makes sense to talk about the Spohn ordinal $\kappa_{\mathcal{O}}(P)$ of a proposition in a given system of spheres: essentially, $\kappa_{\mathcal{O}}(P)$ is the smallest number n such that we can find P-worlds with Spohn ordinal n in $\mathcal{O}$. If there are no P-worlds in $\mathcal{O}$, then we put $\kappa_{\mathcal{O}}(P) = 0$. Formally:

$$\kappa_{\mathcal{O}}(P) := \begin{cases} \min\{\kappa_{\mathcal{O}}(w) \mid w \in P \cap \mathrm{S}\} & P \cap \mathrm{S} \neq \varnothing, \\ 0 & \text{otherwise.} \end{cases}$$

While we can understand Spohn ordinals of possible worlds in given orders to give a numerical measure of the plausibility of a given world (with the number 1 corresponding to maximal plausibility), Spohn ordinals of propositions in given order capture a numerical measure of the plausibility of a proposition.

Given the one-to-one correspondence between systems of spheres and plausibility orders noted above, it makes sense to directly speak of Spohn ordinals of worlds *in plausibility orders*, putting $\kappa_{\mathcal{S}}(w) := \kappa_{\mathrm{sph}(\mathcal{S})}(w)$. Analogously, we put $\kappa_{\mathcal{S}}(P) := \kappa_{\mathrm{sph}(\mathcal{S})}(P)$.

Notice that a plausibility order $\mathcal{S}$ may be uniquely specified by determining its domain S and the Spohn ordinal $\kappa_{\mathcal{S}}(w)$ of each world $w \in \mathrm{S}$. When proceding in this way, it is understood that $w \leq_{\mathcal{S}} v$ iff $\kappa_{\mathcal{S}}(w) \leq \kappa_{\mathcal{S}}(v)$.

1.2. Propositional Attitudes

We routinely ascribe doxastic propositional attitudes to agents, by saying, for example, that an agent believes, knows, or doubts certain things. Such ascriptions tell us something about the opinion of the agent in question about those things. Leaving the agent entertaining the attitude implicit, we can model propositional attitudes as families of doxastic propositions, in the following manner.

1.2.1. Doxastic Propositions. A *doxastic proposition* $\mathbf{P}$ *(on W)* is a function

$$\mathcal{S} \overset{\mathbf{P}}{\longmapsto} \mathbf{P}(\mathcal{S})$$

that assigns a proposition $\mathbf{P}(\mathcal{S}) \subseteq \mathrm{S}$ to each plausibility order $\mathcal{S}$ (on W). We call $\mathbf{P}(\mathcal{S})$ the *proposition denoted by* $\mathbf{P}$ in $\mathcal{S}$.

So a doxastic proposition $\mathbf{P}$ gives us a proposition P for each plausibility order $\mathcal{S}$. In this way, doxastic propositions can be seen as "intensionalized propositions", since just which proposition $\mathbf{P}$ gives us may depend on $\mathcal{S}$, and thus on the epistemic state of our agent.

Our main use for the concept of a doxastic proposition is in stating the next definition.

1.2.2. Propositional Attitudes. A *(doxastic) propositional attitude* is an indexed family of doxastic propositions

$$A := \{AP\}_{P \subseteq W},$$

with the doxastic propositions AP indexed by arbitrary propositions $P \subseteq W$ and satisfying

$$AP(\mathcal{S}) = A(P \cap \mathrm{S})(\mathcal{S}). \quad \textit{(Conservativity)}$$

The conservativity constraint imposed on propositional attitudes derives from the intuition that evaluation in a given structure should not depend on objects

that are not part of that structure, a rather minimal well-behavedness assumption. Note that conservativity will be automatically satisfied if a modal-logical language is used in the usual way, since the denotation of a sentence in a given structure is customarily given by a *subset* of the domain of that structure.[4]

A propositional attitude can equivalently be seen as a function

$$P \overset{A}{\longmapsto} AP$$

taking as input a proposition P, and returning as output a doxastic proposition AP, i.e., another function, from plausibility orders to propositions.

The intuitive way to "read" a given propositional attitude A is as follows: given an order $\mathcal{S}$, and a proposition P, the proposition $AP(\mathcal{S})$ collects *the worlds in* S *in which the agent has the propositional attitude* A *towards* P.

1.2.3. Notation. Given a propositional attitude A, a proposition P, and a plausibility order $\mathcal{S}$, we put

$$A_{\mathcal{S}}P := AP(\mathcal{S}).$$

We also frequently write $\mathcal{S}, w \vDash AP$ to mean that $w \in A_{\mathcal{S}}P$, using the notation familar from modal logic; and we write $\mathcal{S} \vDash AP$ to mean that for any $w \in$ S: $\mathcal{S}, w \vDash AP$.

1.2.4. Irrevocable Knowledge and Simple Belief. Using the concept of a propositional attitude, we can give more formal content to the notions of "hard information" and "soft information" we have introduced above. As we have said, an agent gains hard information by excluding possible worlds from consideration. One can think of hard information as what the agent *(irrevocably) knows*. On the other hand, soft information allows the agent to hierarchize the possible worlds that she still considers as candidates for being the actual world. Of special interest are the worlds she considers *most plausible*, and one can think of the most plausible worlds as what the agent *(simply) believes* in view of her soft information.

Here, the qualifiers "irrevocably" and "simply" are used to distinguish these concepts of knowledge and belief from other, similar notions that we

[4]Our terminology is borrowed from the theory of generalized quantifiers: in this literature, a type $< 1, 1 >$ quantifier Q is called *conservative* if $Q(A, B) = Q(A, A \cap B)$, i.e., given sets A and B, the denotation of Q only depends on the A-part of B.[5] In the present context, the domain S plays roughly the role of A, while the propositional argument P plays the role of B, and the content of our requirement is that the proposition determined by the attitude should only depend on the S-part of P, i.e., $S \cap P$.

will make use of. Let us now define irrecovable knowledge and simple belief formally, using the concept of a propositional attitude:

— *(Irrevocable) knowledge*, denoted by K, is the propositional attitude defined by

$$K_{\mathcal{S}}P := \{w \in \mathrm{S} \mid \mathrm{S} \subseteq P\},$$

i.e., the agent (irrecovably) knows P iff P is satisfied in all worlds in S.

— *(Simple) belief*, denoted by B, is the propositional attitude defined by

$$B_{\mathcal{S}}P := \{w \in \mathrm{S} \mid \mathrm{best}\,\mathcal{S} \subseteq P\},$$

i.e., the agent (simply) believes that P iff P is satisfied in all the most plausible worlds in $\mathcal{S}$.

Irrevocable knowledge is the strongest reasonable formalization of the (pre-formal) concept of knowledge that our setting gives rise to; simple belief, on the other hand, is the weakest reasonable formalization of the (pre-formal) concept of belief that our setting gives rise to.

1.2.5. Introspectiveness. Irrecovable knowledge and simple belief are both examples of *introspective* propositional attitudes.

We call a propositional attitude A *introspective* if, for any proposition P, and for any plausibility order $\mathcal{S}$:

$$A_{\mathcal{S}}P \in \{\mathrm{S}, \varnothing\}.$$

In other words, introspective attitudes are *global properties* of a given plausibility order: either the agent has the attitude A towards P in *all* worlds in S, or in no world in S.

Notice that K as well as B are indeed introspective propositional attitudes. Whether the agent (irrevocably) knows that P in a given order $\mathcal{S}$ does not depend on any particular world in S; it merely depends on the question whether *all* worlds in S are contained in P. Analogously: whether the agent (simply) believes that P in $\mathcal{S}$ does not depend on any particular world in S; it merely depends on the question whether *all* world in best $\mathcal{S}$ are contained in P.

To motivate why we call this property "introspectiveness", let us introduce a number of operations on propositional attitudes that will be useful in the remainder of this dissertation.

1.2.6. Operations on Propositional Attitudes. Given propositional attitudes A and A', one easily defines new ones by means of standard operations. In particular,

— the *opposite* $A^\neg$ of A is given by $A^\neg_{\mathcal{S}}P := A_{\mathcal{S}}(\neg P)$;
— the *complement* $\neg A$ of A is given by $(\neg A)_{\mathcal{S}}P := \mathrm{S} \setminus A_{\mathcal{S}}P$;
— the *dual* $A^\sim$ of A is given by $A^\sim_{\mathcal{S}}P := \neg A^\neg_{\mathcal{S}}P$;
— the *composition* AA' of A and A' is given by $(AA')_{\mathcal{S}}P := A_{\mathcal{S}}(A'_{\mathcal{S}}P)$;
— the *conjunction* $A \wedge A'$ of A and A' is given by $(A \wedge A')_{\mathcal{S}}P := A_{\mathcal{S}}P \cap A'_{\mathcal{S}}P$;
— the *disjunction* $A \vee A'$ of A and A' is given by $(A \vee A')_{\mathcal{S}}P := A_{\mathcal{S}}P \cup A'_{\mathcal{S}}P$;
— the *material implication* $A \to A'$ of A and A' is given by $(A \to A')_{\mathcal{S}}P = (A^\neg \vee (A \wedge A'))_{\mathcal{S}}P$.

We use the notation familiar from modal logic, writing $\mathcal{S} \vDash AP \wedge A'P$ iff $\mathcal{S} \vDash (A \wedge A')P$, or $\mathcal{S} \vDash AP \to A'P$ iff $\mathcal{S} \vDash (A \to A')P$ etc.

Notice now that a propositional attitude A is introspective (cf. §1.2.5 above) iff for any plausibility order $\mathcal{S}$ and proposition P,

$$\mathcal{S} \vDash AP \to KAP$$

and

$$\mathcal{S} \vDash \neg AP \to K\neg AP$$

are both satisfied, that is: if the agent has the attitude A towards P in $\mathcal{S}$, then the agent knows that she has the attitude A towards P ("positive introspection"), and if the agent does not have the attitude A towards P in $\mathcal{S}$, then the agent knows that she does not have the attitude A towards P ("negative introspection"). And this is just how introspectiveness (understod as the conjunction of positive and negative introspection) is usually defined.

An introspective propositional attitude A may be defined by specifying, for each proposition P, the set of plausibility orders $\mathcal{S}$ such that $\mathcal{S} \vDash AP$. This is sufficient since, by definition, $\mathcal{S} \vDash AP$ iff $A_{\mathcal{S}}P = \mathrm{S}$, and by the fact that A is introspective, $\mathcal{S} \nvDash AP$ iff $A_{\mathcal{S}}P = \varnothing$. Hence, for an introspective propositional attitude A, a specification of the plausibility orders $\mathcal{S}$ such that $\mathcal{S} \vDash AP$ fixes, for each plausibility order $\mathcal{S}$ and world $w \in \mathrm{S}$, whether $\mathcal{S}, w \vDash AP$.

1.2.7. Conditional Belief. For any proposition Q, *belief conditional on* Q, denoted by B^Q, is the introspective propositional attitude defined, for each proposition P, by

$$\mathcal{S} \vDash B^Q P \text{ iff } \mathrm{best}_{\mathcal{S}}\, Q \subseteq P,$$

i.e., P is believed conditional on Q iff P is satisfied in all the most plausible Q-worlds in $\mathcal{S}$.

Conditional belief may equivalently be defined in terms of the notion of conditionalization defined above (cf. §1.1.3), noticing that

$$\mathcal{S} \vDash B^Q P \text{ iff } \mathcal{S}|_Q \vDash BP.$$

So P is believed conditional on Q iff P is believed in $\mathcal{S}|_Q$, the conditionalization of $\mathcal{S}$ on Q. This clearly brings out the dynamic flavour of conditional belief: conditional beliefs are beliefs that are held conditional on various pieces of hard information. That is, we can interpret the fact that $\mathcal{S} \vDash B^Q P$ as saying that given the hard information that Q is the case, the agent would believe that P is the case.

Following up on this idea, conditional beliefs provide a way of gauging the "stability" of given propositional attitudes under the influx of hard information. As an example, consider simple belief and irrevocable knowledge, the two propositional attitudes we have defined in §1.2.4. Here, we observe:

PROPOSITION 1 (Baltag and Smets (2008)). *Let $\mathcal{S}$ be a plausibility order, and let P be a proposition. Suppose that $\mathcal{S} \vDash BP$. Then*

— *$\mathcal{S} \vDash BP$ iff $\mathcal{S} \vDash B^S P$.*
— *$\mathcal{S} \vDash KP$ iff $\mathcal{S} \vDash B^Q P$ for any $Q \subseteq W$.*

As it turns out, then, simple beliefs are only guaranteed to be stable when receiving hard information that the agent already has. In this sense, simple beliefs are *easily defeated*. Irrecovable knowledge, on the other hand, are those beliefs that are stable under receiving *any* new hard information. In this sense, irrecovable knowledge is *indefeasible*.

1.2.8. FURTHER EXAMPLES. We now introduce a number of further important examples of propositional attitudes that will be relevant in the remainder of this dissertation.

— *Strong belief*, denoted by Sb, is the introspective propositional attitude defined by

$$\mathcal{S} \vDash SbP \text{ iff } \operatorname{best}\mathcal{S} \subseteq P \text{ and } \forall x \in P \forall y \notin P : x < y,$$

i.e., P is strongly believed iff P is believed, and moreover, all P-worlds are strictly more plausible than all non-P-worlds.[6]

[6] Strong belief is considered by Stalnaker (1996)—who calls the notion "robust belief"—, and by Battigalli and Siniscalchi (2002).

— *Refined belief*, denoted by Rb, is the propositional attitude defined by

$$\mathcal{S} \models RbP \text{ iff } \text{best}\mathcal{S} \subseteq P \text{ and } \forall x \in P, y \notin P : x < y \text{ or } y < x\},$$

i.e., P is refined believed iff P is believed, and moreover, there are no ties between P-worlds and non-P-worlds in the ranking given by $\leq$.[7]

— *Belief to degree n*, denoted by B^n, is the introspective propositional attitude defined, for any natural number $n \geq 1$, by

$$\mathcal{S} \models B^n P \text{ iff } \forall w \in \mathrm{S} : \kappa_{\mathcal{S}}(w) \leq n \Rightarrow w \in P.$$

Notice that B^1 is just B, i.e., $B^1 P$ is satisfied in a given plausibility order $\mathcal{S}$ iff the most plausible worlds in $\mathcal{S}$ are P-worlds. More generally, we have $\mathcal{S} \models B^n P$ iff all worlds with Spohn ordinal from 1 to n are P-worlds.[8]

— *Defeasible knowledge*, denoted by $\Box$, is the propositional attitude defined by

$$\mathcal{S}, w \models \Box P \text{ iff } \forall v \in \mathrm{S} : v \leq w \Rightarrow v \in P,$$

i.e., P is defeasibly known at w iff P is satisfied in all worlds that are at least as plausible as w. This formalizes a notion of knowledge that is weaker than irrevocable knowledge, but still *veridical* in the sense that from that fact that $\mathcal{S}, w \models \Box P$ it follows that $w \in P$.[9]

Observe that defeasible knowledge is not introspective (in the sense of the definition given in §1.2.5): different propositions may be defeasibly known at distinct worlds in a given plausibility order.

All propositional attitudes in the above list allow for natural characterizations in terms of conditional belief. The first and the fourth item of the next proposition are due to Baltag and Smets (2008).

PROPOSITION 2. *Let $\mathcal{S}$ be a plausibility order, and let P be a proposition. Suppose that $\mathcal{S} \models BP$. Then*

[7] To the best of my knowledge, refined belief has not been considered in the previous literature. However, as we shall see later on, refined belief is closely connected to a dynamic attitude called *moderate trust* (defined in §2.5.1), which appears for the first time (under the name of "restrained revision") in the work of Booth and Meyer (2006).

[8] Degrees of belief originate in Spohn's work on ranking functions, cf. Spohn (1988), Aucher (2003), van Ditmarsch (2005). Spohn labels the degrees starting with 0, rather than starting with 1.

[9] Defeasible knowledge was defined by Stalnaker (2006) in his formalization of Lehrer's defeasibility theory of knowledge (Lehrer 1990). Stalnaker defined defeasible knowledge in terms of conditional belief (i.e., as in item (4.) of Proposition 2 below); that defeasible knowledge is simply the Kripke modality for the converse $\geq$ of the plausibility order $\leq$ was discovered by Baltag and Smets (2008), who initially called the notion "safe belief."

1. $\mathcal{S} \vDash SbP$ *iff* $\mathcal{S} \vDash B^{Q}P$ *for any* Q *such that* $Q \cap P \neq \varnothing$.

2. $\mathcal{S} \vDash RbP$ *iff* $\mathcal{S} \vDash B^{Q}P$ *for any* Q *such that* $\operatorname{best} Q \cap P \neq \varnothing$.

3. $\mathcal{S} \vDash B^{n}P$ *iff* $\mathcal{S} \vDash B^{Q}P$ *for any* Q *such that* $Q = \mathrm{S} \setminus \{w \in \mathrm{S} \mid \kappa_{\mathcal{S}}(w) \leq k\}$ *for some natural number* $k < n$.

4. $\mathcal{S}, w \vDash \Box P$ *iff* $\mathcal{S}, w \vDash B^{Q}P$ *for any* Q *such that* $w \in Q$.

PROOF. We prove the second and third item. Start with (2.). From left to right, suppose that $\mathcal{S} \vDash RbP$. Take any Q such that $\operatorname{best} Q \cap P \neq \varnothing$. Then $\mathcal{S} \vDash B^{Q}P$ iff $\operatorname{best} Q \subseteq P$. Take any $w \in \operatorname{best} Q$ and suppose $w \notin P$. Since $\operatorname{best} Q \cap P \neq \varnothing$, there exists v such that $w \approx v$, $v \notin P$. Thus, $\mathcal{S} \nvDash RbP$, contradiction. So $\mathcal{S} \vDash B^{Q}P$. This finishes the left to right direction.

From right to left, suppose that $\mathcal{S} \vDash BP$ and for any Q such that $\operatorname{best} Q \cap P \neq \varnothing$: $\mathcal{S} \vDash B^{Q}P$. If $\mathcal{S} = \varnothing$, our claim holds. If $\mathcal{S} \neq \varnothing$, then $\operatorname{best}\mathcal{S} \neq \varnothing$, so $P \cap \mathrm{S} \neq \varnothing$. Take any $w \in P \cap \mathrm{S}$. Suppose there exists $v \in \neg P \cap \mathrm{S}$ such that $w \approx v$. Then $\operatorname{best}\{w, v\} \cap P \neq \varnothing$ and $\mathcal{S} \nvDash B^{\{w,v\}}P$. This is a contradiction. So for any $v \in \neg P \cap \mathrm{S}$: $w < v$ or $v < w$. It follows that $\mathcal{S} \vDash RP$. This finishes the right to left direction, and the proof for the second item.

We continue with the proof for the third item. From left to right, suppose that $\mathcal{S} \vDash B^{n}P$ for some $n \geq 1$. Take any Q such that $Q = \mathrm{S} \setminus \{w \in \mathrm{S} \mid \kappa_{\mathcal{S}}(w) \leq k\}$ for some natural number $k < n$. Suppose towards a contradiction that $\mathcal{S} \nvDash B^{Q}P$, i.e., it is not the case that $\operatorname{best}_{\mathcal{S}} Q \subseteq P$, and thus there exists a world $v \in \operatorname{best}_{\mathcal{S}} Q$ such that $v \notin P$. Hence, by definition of Q there exists v such that $\kappa_{\mathcal{S}}(v) \leq n$ and $v \notin P$. It follows that $\mathcal{S} \nvDash B^{n}P$. This is a contradiction. Thus $\mathcal{S} \vDash B^{Q}P$ after all, which completes the left to right direction.

From right to left, we argue by contraposition. Suppose that $\mathcal{S} \nvDash B^{n}P$. Then there exists some world $w \in \mathrm{S}$ such that $\kappa_{\mathcal{S}}(w) \leq n$ and $w \notin P$. Hence $\operatorname{best}_{\mathcal{S}} Q \nsubseteq P$, where $Q = (\mathrm{S} \setminus \{w \in \mathrm{S} \mid \kappa_{\mathcal{S}}(w) \leq n-1\})$. So $\mathcal{S} \nvDash B^{Q}P$. Hence it is not the case that $\mathcal{S} \vDash B^{Q}P$ for any Q such that $Q = \mathrm{S} \setminus \{w \in \mathrm{S} \mid \kappa_{\mathcal{S}}(w) \leq k\}$ for some natural number $k < n$. This finishes the right to left direction, and the proof for the third item. ⊣

Let me comment on the individual items of the previous proposition in turn. Throughout, we assume as given a plausibility order $\mathcal{S}$ and a proposition P such that $\mathcal{S} \vDash BP$. Item (1.) says that strong beliefs are particularly *robust*: P is strongly believed in $\mathcal{S}$ iff it is simply believed conditional on any Q consistent with P. Item (2.) says that refined beliefs are robust in a weaker sense. We can say that a proposition P is "not implausible given Q" iff $P \cap \operatorname{best} Q \neq \varnothing$. Then, according to item (2.), "refined belief in P" is the same as "belief in P conditional on any Q such that P is not implausible given Q".

Item (3.) says that P is believed to degree n in $\mathcal{S}$ iff conditionalizing on the set of worlds with Spohn ordinal less than or equal to k yields an order in which P is believed, for any k smaller than n. Item (4.), finally, says that P is defeasibly known at a world w in $\mathcal{S}$ iff the belief in P can be defeated only by receiving hard information that is actually *false* at w, i.e., P is defeasibly known at world w iff P is believed conditional on any proposition Q satisfied at w.

1.2.9. Triviality and Absurdity. Two further special examples of introspective propositional attitudes are given by *triviality* $\top$ and *absurdity* $\bot$, defined by

$$\mathcal{S} \vDash \top P \text{ iff } P \subseteq W,$$

$$\mathcal{S} \vDash \bot P \text{ iff } S = \varnothing.$$

Observe that $\top_{\mathcal{S}} P = KW$ and $\bot_{\mathcal{S}} P = K\varnothing$ for any $\mathcal{S}$ and P. Triviality is the propositional attitude the agent has in *any* possible world of any order; absurdity is the propositional attitude the agent has in *no* possible world of any order.

1.2.10. Duals. Recall from above (§1.2.6) that the *dual* of a propositional attitude A is given by $A^{\sim}_{\mathcal{S}} P := \neg A^{\neg}_{\mathcal{S}} P$. By way of illustration, we give explicit clauses for four examples of propositional attitudes introduced so far below:

— The dual $K^{\sim}$ of irrevocable knowledge K is given by

$$\mathcal{S} \vDash K^{\sim} P \text{ iff } P \cap S \neq \varnothing.$$

— The dual $Sb^{\sim}$ of strong belief Sb is given by

$$\mathcal{S} \vDash Sb^{\sim} P \text{ iff } \exists w, v \in S : w \in P, v \notin P \text{ and } v \leq w.$$

— The dual $B^{\sim}$ of simple belief is given by

$$\mathcal{S} \vDash B^{\sim} P \text{ iff } \operatorname{best}_{\mathcal{S}} \cap P \neq \varnothing.$$

— The dual $\Diamond := \Box^{\sim}$ of defeasible knowledge $\Box$ is given by

$$\mathcal{S}, w \vDash \Diamond P \text{ iff } \exists v : v \leq w \text{ and } v \in P.$$

$K^{\sim}$ expresses the *possibility* of P (i.e., it is not the case that the agent has hard information that not P); $Sb^{\sim}$ expresses the *remote plausibility* of P (i.e., it is not the case that all non-P-worlds are strictly more plausible than all P-worlds); $B^{\sim}$ expresses the *plausibility* of P (i.e., it is not the case that all $\leq$-minimal elements are non-P-worlds); finally, $\Diamond$ expresses a "defeasible possibility", that is, $\mathcal{S}, w \vDash \Diamond P$ iff there is a world v that is at least as plausible as w such that v satisfies P.

1.3. Doxastic Upgrades

As time passes, an agent's epistemic state may change in various ways. While plausibility orders do not track time explicitly, we can model changes over time in a step-wise manner, by changing a given order and moving to another one. For this purpose, we use the notion of an upgrade, i.e., a specific type of relational transformer, that takes given plausibility orders $\mathcal{S}$ to new ones, by (possibly) restricting the domain S, and (possibly) reordering worlds in the relation $\leq$.

1.3.1. Doxastic Upgrades. A *(doxastic) upgrade u on W* is a function

$$\mathcal{S} \overset{u}{\longmapsto} \mathcal{S}^u$$

that takes a given plausibility order $\mathcal{S} = (S, \leq_{\mathcal{S}})$ (on W) to a plausibility order $\mathcal{S}^u = (S^u, \leq_{\mathcal{S}^u})$ (on W), satisfying $S^u \subseteq S$.

The requirement that $S^u \subseteq S$ means that upgrades either grow the hard information of an agent (by eliminating worlds) or leave it the same (by not eliminating worlds). This embodies our understanding that hard information is really *hard*: it does not get lost as the agent's information state changes (hence the "up" in "upgrade"). But crucially, upgrades may *reorder* the plausibility hierarchy (given by the relation $\leq_{\mathcal{S}}$) that represents the agent's soft information (hence the "grade" in upgrade).

1.3.2. Executability. Given a plausibility order $\mathcal{S}$ and a world $w \in S$, an upgrade u is *executable in w* iff $w \in \mathcal{S}^u$. An upgrade is *executable* in $\mathcal{S}$ if it is executable in some $w \in S$, i.e., iff $\mathcal{S}^u \neq \varnothing$.

1.3.3. Examples. Three examples of upgrades have received particular attention in the recent literature: $!P$, $\Uparrow P$ and $\uparrow P$.[10] We add two obviously interesting "edge cases": $\varnothing$ and *id*. So we have five first examples, defined as follows. For each proposition P,

— the *update* $!P$ maps each plausibility order $\mathcal{S}$ to the conditionalization of the order on P, i.e., we simply put $\mathcal{S}^{!P} := \mathcal{S}|_P$: all non-$P$-worlds are deleted, while the new order on the remaining worlds is inherited from $\mathcal{S}$.[11]

[10] Cf. in particular the papers by van Benthem (2007) and Baltag and Smets (2008), and the monograph by van Benthem (2011).

[11] This notion seems to be folklore. It is implicit in Stalnaker (1978) and appears frequently in the dynamic semantics literature (cf. van Eijck and Visser (2008) for an overview). The present notion of update may also be seen as the natural qualitative analogue of Bayesian update.

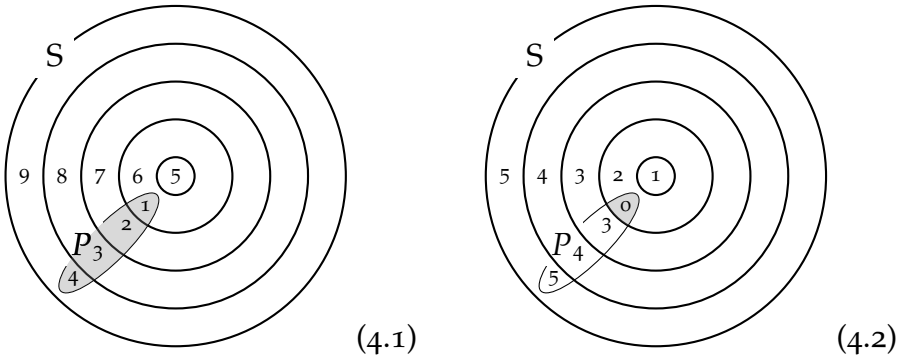


FIGURE 4. Lexicographic Upgrade $\Uparrow P$ (4.1) and Minimal Upgrade $\uparrow P$ (4.2). Applying $\Uparrow P$ amounts to putting all P-worlds on top of all non-P-worlds, i.e., after the upgrade, all P-worlds are closer to the center than all non-P-worlds (while the order within the two zones remains unchanged). Applying $\uparrow P$ amounts to putting only the best P-worlds on top of all the non-P-worlds, leaving everything else unchanged.

— the *lexicographic upgrade* $\Uparrow P$ makes all P-worlds strictly better than all non-P-worlds. No worlds are deleted, and in-between the two "zones" P and $\neg P$, the order is not affected by the upgrade (see Figure 4.1 for illustration).[12]
— the *minimal upgrade* $\uparrow P$ makes the best P-worlds the best worlds overall. No worlds are deleted, and the relation among all worlds that are not in $\text{best}_{\mathcal{S}} P$ remains the same (see Figure 4.2 for illustration).[13]
— the *null upgrade* $\varnothing$ maps every plausibility order to the empty plausibility order.
— the *trivial upgrade id* maps every plausibility order to itself.

1.3.4. COMPOSITION. Upgrades are functions, so they can be composed to obtain new functions. Given arbitrary upgrades u and u', the *composition* $u \cdot u'$ *of* u *and* u' is given by $\mathcal{S}^{u \cdot u'} := (\mathcal{S}^u)^{u'}$.

PROPOSITION 3. *For any upgrade* u*:*

1. $u \cdot id = id \cdot u = u$.

2. $u \cdot \varnothing = \varnothing \cdot u = \varnothing$.

[12] Cf. Nayak (1994).
[13] Cf. Boutilier (1996).

PROOF. The first item is obvious. The second item follows from the fact that $\varnothing^u = \varnothing$ by definition of upgrades. ⊣

1.3.5. SUBSUMPTION. Let $\mathcal{S}$ be a plausibility order, and let u be an upgrade. We say that $\mathcal{S}$ *subsumes* u (or: $\mathcal{S}$ is *in the fixed point of* u) iff $\mathcal{S}^u = \mathcal{S}$.

This captures that the information carried by the upgrade u is "already present" in the order $\mathcal{S}$: actually applying the upgrade is thus redundant.

This naturally gives rise to a notion of relative subsumption of upgrades: an upgrade u is said to subsume an upgrade u' if applying u generally makes applying u' redundant.

Formally, let u, u' be upgrades. We say that u *subsumes* u' iff applying u generally yields an order in the fixed point of u'; that is, for any plausibility order $\mathcal{S}$:

$$(\mathcal{S}^u)^{u'} = \mathcal{S}^u.$$

We write $u \vDash u'$ if u subsumes u'.

Two examples: one may easily check that $!P$ subsumes $\Uparrow P$, which in turn subsumes $\uparrow P$, for any proposition P.

The concept of subsumption is familiar from the dynamic semantics literature. There, the analogue of subsumption of an upgrade by an order is often called *support*, or *acceptance*, and the analogue of subsumption of an upgrade by another is called, simply, *dynamic entailment*.[14]

1.4. Uptake Operators

Our examples of upgrades naturally suggest to isolate the propositional argument already implicit in the definition of concrete upgrades like $!P$ or $\uparrow P$. This leads to the notion of an "uptake operator".

1.4.1. UPTAKE OPERATORS. An *(uptake) operator* τ *(on* W*)* is a family of upgrades

$$\{\tau P\}_{P \subseteq W},$$

[14] Cf. Veltman (1996). The notion of ("relative") subsumption discussed here corresponds to Veltman's entailment relation "validity$_2$": "an argument is valid$_2$ iff updating any information state σ with the premises $\psi_1, \ldots, \psi_n$ in that order, yields an information state in which the conclusion φ is accepted." (ibid.) In the simplest case we have just a single premise, and then, φ dynamically entails ψ (in the sense of validity$_2$) if updating with φ yields a state in which ψ is accepted. "Acceptance", in turn, is defined in terms of fixed points: "every now and then it may happen that $\sigma[\varphi] = \sigma$. If so, the information conveyed by φ is already subsumed by σ. In such a case we (...) say that φ is accepted in σ." (ibid.) So Veltman's notion of acceptance corresponds to our notion of a state subsuming an upgrade.

indexed by arbitrary propositions $P \subseteq W$. Equivalently, uptake operators can be defined as functions from propositions to upgrades, i.e., τ can be seen as a function

$$P \overset{\tau}{\longmapsto} \tau P$$

associating each proposition P with an upgrade τP.

Uptake operators represent modes of processing informational inputs given by propositions. Some but, as we will argue below, not all uptake operators can be seen as describing *strategies for belief change*. Just which uptake operators do correspond to such strategies will be the topic of §1.5.

An uptake operator τ which describes a strategy for belief change, does so in the following way, viewed from the perspective of the agent:

> "Whenever I receive the information that P from a τ-source, I will change my belief state from my current plausibility order $\mathcal{S}$ to $\mathcal{S}^{\tau P}$."

So an uptake operator τ, insofar as it corresponds to a dynamic attitude, captures an assessment of reliability by means of exhaustively describing *behavioural dispositions*: τ encodes, for each possible epistemic state, represented by a plausibility order $\mathcal{S}$, and informational input, represented by a proposition P, how the agent reacts when receiving the input P in state $\mathcal{S}$: by proceding to a new epistemic state, represented by the plausibility order $\mathcal{S}^{\tau P}$.

1.4.2. Infallible, Strong and Minimal Trust. Our three examples of standard upgrades readily give rise to examples of uptake operators.

— *Infallible trust* ! is the operator that maps each proposition P to the update $!P$. The operator ! captures that the source is *known to be infallible*.
— *Strong trust* $\Uparrow$ is the operator that maps each proposition P to the lexicographic upgrade $\Uparrow P$. The operator $\Uparrow$ captures that the source of information is strongly believed (but not known) to be trustworthy.
— *Minimal trust* $\uparrow$ is the operator that maps each proposition P to the minimal upgrade $\uparrow P$. The operator $\uparrow$ captures that the source of information is simply believed (but not known or strongly believed) to be trustworthy.

These operations formalize three distinct levels of trust. It is natural to relate them to Spohn's tiger example discussed in the introduction of this dissertation. Recall our set of scenarios:

— *I read a somewhat sensationalist coverage in the yellow press claiming that there are tigers in the Amazon jungle.*

— *I read a serious article in a serious newspaper claiming this.*
— *I read the Brazilian government officially announcing that tigers have been discovered in the Amazon area.*
— *I see a documentary on TV claiming to show tigers in the Amazon jungle.*
— *I read an article in Nature by a famous zoologist reporting of tigers there.*
— *I travel to the Amazon jungle, and see the tigers.*

Infallible trust is bestowed upon a source that is considered to be an absolute, unquestionable authority concerning the truth of the information received. Infallible trust might thus be applicable to "seeing the tigers with one's own eyes", or perhaps to "reading about the tigers in Nature." But the latter case could also be treated as a case of strong trust, which captures a strong, but still defeasible form of trust in a source—upon reading about the tigers in Nature, one might still want to travel there and check whether the information obtained is correct. A documentary on TV, or an official government announcement on the other hand, might be taken to correspond to minimal trust, belief-inducing, but at the same time, easily defeasible.

1.4.3. Neutrality and Isolation. The trivial upgrade and the null upgrade introduced earlier give rise to two special uptake operators, called "neutrality" and "isolation".

— *Neutrality id* is the operator given by $\mathcal{S}^{idP} := \mathcal{S}$.
— *Isolation* $\varnothing$ is the operator given by $\mathcal{S}^{\varnothing P} := \varnothing$.

The uptake operator *id* formalizes the concept of "ignoring a source": receiving the information that P from an *id*-source is completely inconsequential. Let us consider a "real life" example: Sometimes, information received from a source is best ignored.

Bart to Jessica: *I really am the man of your dreams.*

Jessica dismisses what Bart says. From a certain perspective, it is as if she had not received any information at all; of course, in some way, she will record that Bart uttered that sentence; possibly, that might affect her perception of Bart; etc—but we disregard such aspects here, focusing just on the question how Jessica's opinion *about the proposition in question* is affected by the information obtained from Bart. And the plausible answer is: not at all. More generally, input from a source will be ignored if the source is considered

to be irrelevant; in the narrow sense described, the information received will then not register at all in the epistemic state of the hearer.

The operator $\varnothing$, on the other hand, captures, as the name suggests, isolation from a source: no information received from a $\varnothing$-source will ever give rise to an executable upgrade. This is encoded in the fact that isolation encodes an irreversible crash of an agent's belief system due to an arbitrary piece of information: for any order $\mathcal{S}$, and proposition P, $\mathcal{S}^{\varnothing P} = \varnothing$, and for any proposition Q and operator τ, $\varnothing^{\tau Q} = \varnothing$: no information received from another source will be able to repair the damage done by upgrading with information from a $\varnothing$-source. This dynamic attitude, $\varnothing$, is useful if we want to model that the communication channel from a particular source to our agent is blocked: the agent cannot, actually, receive any information from that source.

1.4.4. Creating Propositional Attitudes. Given an uptake operator τ, it is natural to ask what propositional attitudes are induced by upgrades τP applied to arbitrary plausibility orders.

For an uptake operator τ and a propositional attitude A, we say that τ *creates* A iff $\mathcal{S}^{\tau P} \vDash AP$, for any order $\mathcal{S}$ and proposition P.

proposition 4.

— *Infallible trust* ! *creates knowledge* K.
— *Strong trust* $\Uparrow$ *creates the disjunction of opposite knowledge and strong belief* $K^{\neg} \vee Sb$.
— *Minimal trust* $\uparrow$ *creates the disjunction of opposite knowledge and belief* $K^{\neg} \vee B$.
— *Isolation creates absurdity.*
— *Neutrality creates triviality.*

proof. We do the first two items for illustration. For the first item, let $\mathcal{S}$ be a plausibility order, and P a proposition. Then $\mathrm{S}^{\tau P} = \mathrm{S} \cap P \subseteq P$, hence $\mathcal{S}^{\tau P} \vDash KP$. We conclude that infallible trust creates knowledge. For the second item, let, again, $\mathcal{S}$ be a plausibility order, and P a proposition. If $P \cap \mathrm{S} = \varnothing$, then $P \cap \mathrm{S}^{\tau P} = \varnothing$ since $\mathrm{S}^{\tau P} \subseteq \mathrm{S}$, hence $\mathcal{S}^{\tau P} \vDash K^{\neg} P$. If, on the other hand, $P \cap \mathrm{S} \neq \varnothing$, then by definition of strong trust, $\mathcal{S}^{\tau P} \vDash SbP$. Hence strong trust creates the disjunction of opposite knowledge and strong belief. ⊣

The connection between strong trust and minimal trust on the one hand, and strong belief and simple belief on the other hand is thus, in view of the previous proposition, not as straightforward as one might have expected: in general, strong trust does *not* create strong belief, but rather the disjunction of opposite knowledge and strong belief, and similarly, minimal trust does *not* create simple belief, but rather the disjunction of opposite knowledge and simple belief. We follow up on these observations in §2 and §1.7.

1.5. Dynamic Attitudes

We have seen a number of examples of uptake operators that intuitively correspond to different forms of trust. The question is now the following: *which uptake operators can reasonably be taken to represent dynamic attitudes, understood in a pre-formal sense, as agent-internal assessments of the reliability of a source?*

1.5.1. Framework- vs. Theory-Level Constraints. In the Belief Revision literature, many attempts of formulating theories of what counts as "rational revision" exist. The AGM postulates and the postulates on transformations of plausibility orders aiming at characterizing "rational iterated revision" suggested by Darwiche and Pearl are examples.[15] But the above query is of a different kind. The framework should allow more variety than the theories (for example, theories of rational revision) one may develop in it. The following quote pertains to a different problem domain—the logical study of time—, but clearly brings out the difference:

> *In the logical study of Time, attention is often restricted to the choice of specific axioms for the temporal precedence order matching certain desired validities in the tense logic. But, there [also] (. . .) exist preliminary global intuitions, such as 'anisotropy' or 'homogeneity', constituting the texture of our idea of Time, constraining rather than generating specific relational conditions.*[16]

Our investigation of the concept of a dynamic attitude begins with an attempt to identify a number of "global intuitions" in Van Benthem's sense, i.e., framework-level constraints, embodying the "texture" of our idea of a dynamic attitude, rather than "specific axioms", i.e., theory-level requirements. Such "specific axioms" have historically played the predominant role in the belief revision literature, as the AGM tradition basically started with a list of such constraints, known today as "the AGM postulates." For our purposes, however, it seems methodologically helpful to avoid, as best as we can, encoding substantive requirements at the framework-level.[17]

Another difference of our approach to the AGM tradition is its broader focus. The latter theory was developed with an interest in modeling revision with information obtained from a *reliable* source. Here, we are not only interested in the acceptance of new information (based on trust), but also in

[15] Alchourrón et al. (1985), Darwiche and Pearl (1996).

[16] van Benthem (1984).

[17] Of course, what exactly counts as a framework-level rather than theory-level constraint may itself be a matter of debate.

its rejection (based on distrust), as well as in various intermediate levels, and ways of "mixing" trust and distrust. This provides another reason why the constraints we are looking for need to be found at a higher level of abstraction.

In the remainder of this section, we propose four such framework-level constraints: *idempotence*, *dynamic conservativity*, *informativity* and *closure under opposites*. We specify and motivate them, and use them to define our notion of a dynamic attitude.

1.5.2. Idempotence. An operator τ is *idempotent* iff for any $P \subseteq W$: $\tau P \models \tau P$, where we recall that $\tau P \models \tau P$ iff for any plausibility order $\mathcal{S}$: $(\mathcal{S}^{\tau P})^{\tau P} = \mathcal{S}^{\tau P}$ (cf. §1.3.5).

As a constraint on dynamic attitudes, the condition is motivated by the consideration that receiving the same information (individuated *semantically*, as a proposition) from the same source one has *just received it from* should be redundant. Processing the same information twice is unnecessary. Another way to motivate this is our assumption that dynamic attitudes describe, comprehensively, in totality, how an agent processes information coming from a particular source. We can think of each particular information processing event as moving to a new stable state based on taking the input into account. Then our constraint says that a dynamic attitude τ captures *this stable state*, i.e., $\mathcal{S}^{\tau P}$, for each proposition P and plausibility order $\mathcal{S}$. In other words, τ can be taken to represent the *target of revision* predicated on a particular assessment of reliability.

1.5.3. Dynamic Conservativity. An operator τ *satisfies dynamic conservativity* if for any plausibility order $\mathcal{S}$ and proposition P, $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau(P \cap \mathrm{S})}$.

This requirement parallels the conservativity constraint imposed on propositional attitudes, reflecting the aforementioned fundamental logical principle that properties of a structure do not (cannot, should not) depend on objects that are not part of the structure. In our domain of dynamic attitudes, this corresponds to the idea that worlds that the agent has already irrecovably excluded from consideration (in the sense captured by irrecovable knowledge: w is irrecovably excluded in $\mathcal{S}$ if $\mathcal{S} \models K\neg\{w\}$, i.e., if $w \notin \mathrm{S}$) should not affect how she processes an informational input. Hence the effect of applying an upgrade to an order $\mathcal{S}$ should not depend on (all of) P, but rather only on the S-part of P.

1.5.4. Informativity. An operator τ *satisfies informativity* iff for any order $\mathcal{S}$: $\mathcal{S}^{\tau\varnothing} \in \{\mathcal{S}, \varnothing\}$.

The rationale for this requirement is the intuition that absurd information is not useable: receiving the absurd proposition $\varnothing$ from a source, the agent needs to ignore the information (this corresponds to $\mathcal{S}^{\tau\varnothing} = id$), or else the agent will acquire inconsistent beliefs (this corresponds to $\mathcal{S}^{\tau\varnothing} = \varnothing$). Put conversely, if a given input P *does* provide genuine information to an agent (i.e., $\mathcal{S}^{\tau P} \notin \{\mathcal{S}, \varnothing\}$), then $P \neq \varnothing$.

1.5.5. Closure under Opposites. A class of uptake operators Θ is *closed under opposites* if whenever $\tau \in \Theta$, then also $\tau^{\neg} \in \Theta$, where $\tau^{\neg}$ (the *opposite of* τ) is defined by $\tau^{\neg} P := \tau(\neg P)$.[18]

We will require that the class of dynamic attitudes is closed under opposites. The intuition behind this requirement is that corresponding to any particular assessment of reliability (of a source), one should be able to identify a corresponding "assessment of unreliability", which is essentially given by the instruction to "upgrade to the contrary." In effect, this generalizes the idea underlying Smullyan's liar puzzles, where we are effortlessly able to think of sources as "opposite truthtellers", to be treated just in the way formalized by the opposite operation, and based on our understanding of what it means to treat someone as a "truthteller".

1.5.6. Dynamic Attitudes. The preceding considerations lead to the following definition which introduces the central concept of this dissertation.

The class of *dynamic attitudes (on W)* is the largest class Δ of uptake operators τ (on W) satisfying, for any plausibility order $\mathcal{S}$ (on W) and proposition $P \subseteq W$:

— *idempotence*: $\tau P \vDash \tau P$,
— *dynamic conservativity*: $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau(P \cap \mathrm{S})}$,
— *informativity*: $\mathcal{S}^{\tau\varnothing} \in \{\mathcal{S}, \varnothing\}$,
— *closure under opposites*: $\tau^{\neg} \in \Delta$.

Observe that our examples from the previous section satisfy these properties, i.e., if $\tau \in \{!, \Uparrow, \uparrow, id, \varnothing\}$, then $\tau \in \Delta$. We will see many more examples of dynamic attitudes starting in the next chapter.

proposition 5. *Any dynamic attitude τ satisfies the following, for all plausibility orders $\mathcal{S}$ and propositions P:*

1. $\mathcal{S}^{\tau W} \in \{\mathcal{S}, \varnothing\}$.

2. *If* $P \cap \mathrm{S} \in \{\mathrm{S}, \varnothing\}$, *then* $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$.

[18] Observe that the opposite operation is involutive, i.e., $(\tau^{\neg})^{\neg} = \tau$.

PROOF. For the first item, use informativity and closure under opposites. For the second item, use the first item, informativity and dynamic conservativity. ⊣

1.5.7. STRONG INFORMATIVITY. An uptake operator τ *satisfies strong informativity* iff $P \in \{W, \varnothing\}$ implies $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$ for any order $\mathcal{S}$ and $P \subseteq W$.

Strong informativity expresses that the absurd proposition $\varnothing$ as well as the trivial proposition W cannot provide useable information for an agent. By its definition and the previous proposition, any dynamic attitude satisfies strong informativity.

In fact, the class of dynamic attitudes Δ can equivalently be defined as the class of uptake operators Δ' satisfying idempotence, dynamic conservativity and strong informativity.

LEMMA 6. $\Delta = \Delta'$.

PROOF. For the "$\subseteq$" direction, any operator $\tau \in \Delta$ satisfies idempotence and dynamic conservativity by its definition, and strong informativeness by its definition together with the first item of Proposition 5. So $\tau \in \Delta'$.

For the "$\supseteq$" direction, let $\tau \in \Delta'$. By definition, τ satisfies idempotence, dynamic conservativity and informativity. It remains to show that $\tau^\neg \in \Delta'$, which implies that $\tau \in \Delta$. To establish the latter claim, first observe that since $\tau^\neg P = \tau(\neg P)$ for any P, $\tau^\neg$ is idempotent by idempotence of τ. Next, we have to show that $\tau^\neg$ satisfies dynamic conservativity. For this, observe that $\mathcal{S}^{\tau^\neg P} = \mathcal{S}^{\tau(\neg P)}$ and $\mathcal{S}^{\tau^\neg (P \cap S)} = \mathcal{S}^{\tau\neg(P \cap S)}$ by definition. Since τ satisfies dynamic conservativity, $\mathcal{S}^{\tau\neg(P\cap S)} = \mathcal{S}^{\tau\neg(P\cap S)\cap S}$. But $\neg(P \cap S) \cap S = \neg P$, so $\mathcal{S}^{\tau\neg(P\cap S)\cap S} = \mathcal{S}^{\tau(\neg P)}$. Hence $\mathcal{S}^{\tau^\neg P} = \mathcal{S}^{\tau^\neg(P\cap S)} = \mathcal{S}^{\tau(\neg P)}$, so $\tau^\neg$ satisfies dynamic conservativity. Finally, strong informativity is obviously satisfied by $\tau^\neg$ whenever strong informativity is satisfied by τ. So $\tau^\neg \in \Delta'$, thus $\tau \in \Delta$, and the proof is complete. ⊣

1.6. Idempotence and Moorean Phenomena

In this section, we discuss how the idempotence condition we have imposed on dynamic attitudes can be squared with the *failures* of idempotence that have received a lot of attention in the literature, both in dynamic epistemic logic and in dynamic semantics. Roughly, the message of this section is that idempotence in the current setting, and idempotence in the setting of dynamic epistemic logic do not amount to the same thing, because the former notion is formulated in terms of propositions, i.e., semantic objects, while the latter notion is formulated w.r.t. a logical language.

1.6.1. Moore Sentences. Suppose that an infallibly trusted source tells you:

It is raining even though you don't know it.

In a situation where your hard information excludes neither that it is raining nor the opposite, that it is not raining, this may very well be useful information. Such a sentence is called a *Moore sentence*. It has the peculiar property that learning it makes it become false.[19] We may formalize our sentence in the syntax of epistemic logic (formally defined right below) as

$$p \wedge \neg Kp.$$

Suppose that the set of worlds where p is true is given by the proposition P. By our assumption, your plausibility order $\mathcal{S}$ is such that neither $P \cap S$ nor $\neg P \cap S$ are empty.

Upon learning the above sentence, you delete all non-P-worlds from $\mathcal{S}$ (remember our assumption: the source is infallibly trusted). But now suppose that the source tells you the same sentence *again*. This time, you delete *all remaining worlds* (since, by now, you know that P), ending up with the empty plausibility order. Clearly, in this case, processing the same input twice is not the same as processing it once.

Is this at odds with the setting we have introduced so far? It's not. To make very clear that this is a non-problem, let us give the syntax and semantics for a standard logical language, and discuss the issue with more precision. Outside of the current discussion, we will not use this language for quite some time (we return to it only in §5).

1.6.2. The Epistemic Language and Its Semantics. Fix a non-empty set Φ, called the set of *atomic sentences*. The elements of Φ are understood as denoting basic facts that may or or may not hold in a given possible world. A *valuation* $[\![\cdot]\!]$ is a map

$$p \overset{[\![\cdot]\!]}{\longmapsto} P$$

is a map that assigns a proposition $P \subseteq W$ to every atomic sentence $p \in \Phi$. A *(single-agent) plausibility model* is a pair $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$, where $\mathcal{S}$ is a plausibility order, and $[\![\cdot]\!]$ is a valuation.

The language $\mathcal{L}$ (called the *epistemic-doxastic language*) is given by the following grammar ($p \in \Phi$):

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid K\varphi$$

[19] Cf. van Ditmarsch and Kooi (2006), Holliday and Icard (2010) for more background on and technical explorations of "Moorean phenomena" in epistemic logic.

Read $K\varphi$ as *the agent infallibly (*or: *indefeasibly) knows that* φ; read $\Box\varphi$ as *the agent defeasibly knows that* φ.

We interpret the language $\mathcal{L}$ in the usual manner, by providing, for each plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ a map $[\![\cdot]\!]_{\mathcal{M}}$ that assigns a proposition $[\![\varphi]\!]_{\mathcal{M}} \subseteq S$ to each sentence $\varphi \in \mathcal{L}$, the proposition comprising the worlds where φ is *satisfied* in $\mathcal{M}$ (equivalently: those worlds w such that φ is *true at* w in $\mathcal{M}$).

$[\![\cdot]\!]_{\mathcal{M}}$ is defined, for each $\mathcal{M}$, by induction on the construction of φ. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model.

$$\begin{aligned}
[\![p]\!]_{\mathcal{M}} &:= [\![p]\!] \cap S,\\
[\![\neg\varphi]\!]_{\mathcal{M}} &:= S \setminus [\![\varphi]\!]_{\mathcal{M}},\\
[\![\varphi \wedge \psi]\!]_{\mathcal{M}} &:= [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}},\\
[\![\Box\varphi]\!]_{\mathcal{M}} &:= \Box_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}},\\
[\![K\varphi]\!]_{\mathcal{M}} &:= K_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}}.
\end{aligned}$$

1.6.3. Intensional Dynamic Attitudes. We now observe that, given the above semantics, any dynamic attitude τ can be lifted to an *intensional dynamic attitude*, which we again denote with τ, and which is given by the set of upgrades

$$\{\tau\varphi\}_{\varphi\in\mathcal{L}},$$

where for each $\varphi \in \mathcal{L}$, $\tau\varphi$ is determined by putting, for every plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$:

$$\mathcal{M}^{\tau\varphi} := \mathcal{M}^{\tau[\![\varphi]\!]_{\mathcal{M}}},$$

where we use the notation $\mathcal{M}^{\tau P} := (\mathcal{S}^{\tau P}, [\![\cdot]\!])$.

The crucial observation is now that it is *not* in general the case that

$$(\mathcal{M}^{\tau\varphi})^{\tau\varphi} = \mathcal{M}^{\tau\varphi}.$$

Consider the case we started with: put $\tau = !$, consider the sentence $p \wedge \neg Kp$, and pick a plausibility model $\mathcal{M}$ such that neither $[\![p]\!]_{\mathcal{M}}$ nor $[\![\neg p]\!]_{\mathcal{M}}$ are empty. Then, as we have seen,

$$(\mathcal{M}^{!(p\wedge\neg Kp)})^{!(p\wedge\neg KP)} \neq (\mathcal{M}^{!(p\wedge\neg Kp)}).$$

What *is* still the case in general, however, is that for every plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$:

$$(\mathcal{S}^{\tau[\![\varphi]\!]_{\mathcal{M}}})^{\tau[\![\varphi]\!]_{\mathcal{M}}} = \mathcal{S}^{\tau[\![\varphi]\!]_{\mathcal{M}}}.$$

So while the dynamic attitude $! = \{\tau P\}_{P\subseteq W}$ is idempotent (as are *all* dynamic attitudes, by definition), the intensional dynamic attitude $! = \{\tau\varphi\}_{\varphi\in\mathcal{L}}$ is *not* idempotent.

This resolves the above "puzzle": our requirement that dynamic attitudes be idempotent is not at odds with the phenomenon of Moore sentences. While an intensional dynamic attitude $\{\tau\varphi\}_{\varphi\in\mathcal{L}}$, which applies to sentences, need not be idempotent, "ordinary" dynamic attitudes $\{\tau P\}_{P\subseteq W}$ are (by definition) idempotent. Processing the same *sentence* received from the same source may fail to yield the same result as processing the sentence only once. But processing the same *proposition* received from the same source amounts to the same as processing it once.

1.7. Fixed Points of Dynamic Attitudes

As pointed out in the introduction, one of our main interests is to study the connection between propositional and dynamic attitudes. This section and the following one, establish a concrete link that forms the basis for much of the work in the dissertation. In this section, we demonstrate that introspective propositional attitudes can be seen as *fixed points* of dynamic ones; in the next section, we introduce a *subsumption order* on dynamic attitudes that turns out to match the *usual entailment* order on the corresponding fixed points.

1.7.1. Fixed Points. Given a dynamic attitude τ, the *fixed point* $\overline{\tau}$ of τ is the introspective propositional attitude $\overline{\tau} = \{\overline{\tau}P\}_{P\subseteq W}$ defined by

$$\mathcal{S} \vDash \overline{\tau}P \text{ iff } \mathcal{S}^{\tau P} = \mathcal{S}.$$

Given a propositional attitude A and a dynamic attitude τ, if $\overline{\tau} = A$, then we often say that τ *realizes* A.

Notice that fixed points are *unique*: every dynamic attitude has exactly one fixed point. To see this, suppose that A_1 and A_2 are both fixed points of τ. Let $\mathcal{S}$ be a plausibility order, and P a proposition. Then $\mathcal{S} \vDash A_1P$ iff (since A_1 is the fixed point of τ) $\mathcal{S}^{\tau P} = \mathcal{S}$ iff (since A_2 is the fixed point of τ) $\mathcal{S} \vDash A_2P$. Since $\mathcal{S}$ and P were arbitrarily chosen, we conclude that $A_1 = A_2$.

Fixed points of dynamic attitudes τ capture the *redundancy* of specific upgrades given by τ by means of the propositional attitude $\overline{\tau}$.[20] According to our definition, $\mathcal{S} \vDash \overline{\tau}P$ iff $\mathcal{S}$ is in the fixed point of the upgrade τP (hence our choice of terminology). What this means intuitively is that $\mathcal{S} \vDash \overline{\tau}P$ iff the agent *already has* the information she would obtain if a τ-source provided her with the input P. Actually receiving P from such a τ-source is thus redundant; the current plausibility order $\mathcal{S}$ "subsumes" the input τP.

[20] Observe that the fixed point $\overline{\tau}$ of a dynamic attitude τ is indeed a propositional attitude, since the dynamic conservativity of τ ensures the conservativity of $\overline{\tau}$.

1.7.2. Creating and Stopping. Given a dynamic attitude τ and a propositional attitude A, the first is the fixed point of the second iff two things come together. Stating them separately is conceptually helpful. We say that τ *creates* A iff $\mathcal{S}^{\tau P} \models AP$ (for any $\mathcal{S}$ and P); and we say that A *stops* τ iff whenever $\mathcal{S} \models AP$, then $\mathcal{S}^{\tau P} = \mathcal{S}$ (for any $\mathcal{S}$ and P).

proposition 7. *The fixed point of τ is A iff τ creates A and A stops τ.*

proof. From left to right, suppose that $\overline{\tau} = A$. Then obviously A stops τ. Further, take any order $\mathcal{S}$ and proposition P. Then $\mathcal{S}^{\tau P} \models \overline{\tau}P$ by idempotence of dynamic attitudes. By the assumption, $\mathcal{S}^{\tau P} \models AP$. So τ creates A. From right to left, suppose that τ creates A and A stops τ. We need to show that $\overline{\tau} = A$. One half of this follows from the fact that A stops τ. For the other half, suppose that $\mathcal{S}^{\tau P} = \mathcal{S}$. Since τ creates A, $\mathcal{S}^{\tau P} \models AP$, i.e., $\mathcal{S} \models AP$, the desired result. So $\overline{\tau} = A$, and the proof is complete. ⊣

So the fact that A is the fixed point of τ indicates that (1) τ creates A and (2) τ leaves the order unchanged once A has been reached. In this sense, τ "dynamically realizes" $\overline{\tau}$.

Two examples illustrate how creating and stopping may come apart. First, consider infallible trust ! and belief B. On the one hand, ! creates B. On the other hand, from the fact that $\mathcal{S} \models BP$, it does not follow that $\mathcal{S}^{!P} = \mathcal{S}$ (counterexample: assume that there are non-P-worlds in S). In our terminology: B does not stop !. So the fixed point of ! is not B. Second, consider absurdity $\bot$ and minimal trust $\uparrow$. On the one hand, $\bot$ stops $\uparrow$; but, obviously, $\uparrow$ does not create $\bot$. So the fixed point of $\uparrow$ is not $\bot$.

1.7.3. Definability of Propositional Attitudes. Here are five first examples of fixed points:

proposition 8.

1. $\overline{!} = K$ *(the fixed point of infallible trust is knowledge).*
2. $\overline{\Uparrow} = Sb \vee K^{\neg}$ *(the fixed point of strong trust is the disjunction of strong belief and the opposite of knowledge).*
3. $\overline{\uparrow} = B \vee K^{\neg}$ *(the fixed point of minimal trust is the disjunction of simple belief and the opposite of knowledge).*
4. $\overline{\varnothing} = \bot$ *(the fixed point of isolation is inconsistency).*
5. $\overline{id} = \top$ *(the fixed point of neutrality is triviality).*

Notice that, perhaps unexpectedly, strong trust $\Uparrow$ and strong belief Sb on the one hand, and minimal trust $\uparrow$ and simple belief B on the other hand, do not quite match. Take $\Uparrow$. The reason for the mismatch is that a plausibility order $\mathcal{S}$ containing no P-worlds does not satisfy SbP. However, $\mathcal{S}^{\Uparrow P} = \mathcal{S}$ by definition of $\Uparrow$. So the fixed point of strong trust is not strong belief. Analogous remarks apply to $\uparrow$ and B. We will see in §2.1 how to define dynamic attitudes whose fixed points are strong belief and simple belief, respectively.

1.7.4. Characterizing Introspectiveness. Fixed points of dynamic attitudes are, by their definition, *introspective* propositional attitudes. In fact, the class of introspective propositional attitudes can be characterized in terms of dynamic attitudes.

For use in the proof of the next theorem, we define, for any propositional attitude A, the dynamic attitude *test for* A, denoted by $?A$, given by

$$\mathcal{S}^{?AP} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash AP, \\ \varnothing & \text{otherwise,} \end{cases}$$

We now observe the following:

theorem 9. *Let A be a propositional attitude. The following are equivalent:*

1. *A is introspective.*

2. *There exists a dynamic attitude τ such that $\overline{\tau} = A$.*

proof. For the direction from (1.) to (2.), suppose that A is an introspective propositional attitude. Consider the dynamic attitude $?A$, as defined ahead of this proposition. The fixed point of $?A$ is, obviously, A, which finishes one direction. Since the other direction is trivial, we are done. ⊣

Think of introspective propositional attitudes as possible *targets of belief change*. The result ensures that for each such target, there is a strategy, given by a dynamic attitude, that realizes that target. Of course, there may be more *reasonable* strategies than the ones actually chosen in the above proof. We further comment on this aspect below, in §1.8.

1.7.5. More on Tests. The dynamic attitude $?A$ we have defined for each propositional attitude A in the previous paragraph tests, for each order $\mathcal{S}$ and proposition P, whether the agent has the propositional attitude A towards P. If the test succeeds, the order remains unchanged; if the test fails, the agent ends up in the absurd epistemic state given by the empty plausibility order $\varnothing$. Intuitively, such tests correspond to *dynamic acts of introspection* of the agent.

We can also use tests to lift operations on propositional attitudes to the dynamic level. Two examples that will be used in the proof of Proposition 12 in §1.8 below: (1) given dynamic attitudes σ and τ, the *disjunction test* $?(\sigma \vee \tau)$ is the dynamic attitude defined by

$$\mathcal{S}^{?(\sigma\vee\tau)P} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash \overline{\sigma}P \vee \overline{\tau}P \\ \varnothing & \text{otherwise} \end{cases}$$

The fixed point of the disjunction test $?(\sigma \vee \tau)$ is the disjunction of the fixed points of σ and τ. (2) given dynamic attitudes σ and τ, the *conjunction test* $?(\sigma \wedge \tau)$ is the dynamic attitude defined by

$$\mathcal{S}^{?(\sigma\vee\tau)P} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash \overline{\sigma}P \wedge \overline{\tau}P \\ \varnothing & \text{otherwise} \end{cases}$$

The fixed point of the conjunction test $?(\sigma \vee \tau)$ is the conjunction of the fixed points of σ and τ.[21]

1.8. Subsumption

1.8.1. Entailment. Propositional attitudes allow for qualitative comparisons along the natural entailment relation. For example, knowledge K implies belief B, since whenever $\mathcal{S} \vDash KP$, also $\mathcal{S} \vDash BP$. More generally, the *entailment order* on (introspective) propositional attitudes is defined by

$$A \leq A' \quad \Leftrightarrow \quad \forall \mathcal{S} \forall P\colon \quad \mathcal{S} \vDash AP \to A'P,$$

for any propositional attitudes A and A'. If $A \leq A'$, then we say that A *entails* A'.

[21] More generally, given an n-ary operation o that assigns an introspective attitude $o(A_1, \ldots, A_n)$ to given introspective attitudes $A_1, \ldots, A_n$, define the n-ary operation $?o$ that assigns to given dynamic attitudes $\tau_1, \ldots, \tau_n$ the test $?o(\tau_1, \ldots, \tau_n)$, defined by

$$\mathcal{S}^{?o(\tau_1,\ldots,\tau_n)P} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash o(\overline{\tau_1}, \ldots, \overline{\tau_n})P \\ \varnothing & \text{otherwise} \end{cases}$$

Then, one may observe that $\overline{?o(\tau_1, \ldots, \tau_n)} = o(\overline{\tau_1}, \ldots, \overline{\tau_n})$. Figure 5 shows this in a diagram. Arrows going down indicate applications of the fixed point operation $\tau \mapsto \overline{\tau}$, while arrows going from left to right indicate applications of $?o$ (at the top), respectively o (at the bottom).

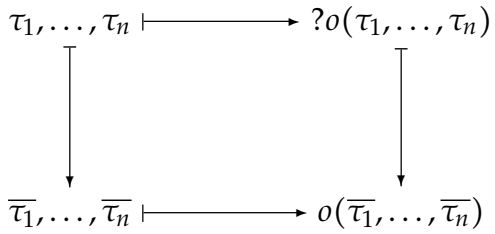


FIGURE 5. In this diagram, downward-point arrows represent applications of the fixed point operator. Taking the fixed point of a bunch of dynamic attitudes $\tau_1, \ldots, \tau_n$ and applying the operation o amounts to the same thing as applying the operation $?o$ to $\tau_1, \ldots, \tau_n$ (as defined in the main text), and only then taking the fixed point of the resulting dynamic attitude $?o(\tau_1, \ldots, \tau_n)$.

1.8.2. SUBSUMPTION. Here, we consider the question how an analogous notion for *dynamic* attitudes should be defined. Our answer is that a strength order on dynamic attitudes naturally arises from the notion of upgrade subsumption defined in §1. Recall that, given upgrades u and u', u subsumes u' (notation: $u \vDash u'$) iff $(\mathcal{S}^u)^{u'} = \mathcal{S}^u$ for any plausibility order $\mathcal{S}$.

We now define the *subsumption order* $\leq$ on dynamic attitudes by putting

$$\sigma \leq \tau \quad \Leftrightarrow \quad \forall P : \sigma P \vDash \tau P$$

for any dynamic attitudes σ and τ. If $\sigma \leq \tau$, then we say that σ *subsumes* τ. We write $\sigma < \tau$ iff $\sigma \leq \tau$ and not $\tau \leq \sigma$; and we write $\sigma \approx \tau$ iff both $\sigma \leq \tau$ and $\tau \leq \sigma$.

1.8.3. EXAMPLES. Here are two examples of subsumption relations among attitudes:

PROPOSITION 10.

— *For any dynamic attitude τ: if $\tau \neq id$, then $\tau < id$, and if $\tau \neq \varnothing$, then $\varnothing < id$.*
— $! < \Uparrow < \uparrow$.

We will see more uses of the subsumption order in §2.5.

1.8.4. SUBSUMPTION AS INCLUSION OF FIXED POINTS. Equivalently, the subsumption order on dynamic attitudes can be defined in terms of fixed points of dynamic attitudes, since an attitude σ subsumes an attitude τ iff the fixed point of σ entails the fixed point of τ. The next theorem justifies this remark.

THEOREM 11. *$\sigma \leq \tau$ iff $\overline{\sigma} \leq \overline{\tau}$.*

PROOF. From (1.) to (2.), suppose that $\sigma \leq \tau$, i.e., for all $\mathcal{S}$, for all P: $(\mathcal{S}^{\sigma P})^{\tau P} = \mathcal{S}^{\sigma P}$. Choose a plausibility order $\mathcal{S}$ and a proposition P, and suppose that $\mathcal{S} \vDash \overline{\sigma} P$. This means that $\mathcal{S}^{\sigma P} = \mathcal{S}$. By the assumption, $(\mathcal{S}^{\sigma P})^{\tau P} = \mathcal{S}^{\sigma P}$, so $\mathcal{S}^{\tau P} = \mathcal{S}$, thus $\mathcal{S} \vDash \overline{\tau} P$, which concludes the left to right direction.

Conversely, suppose that $\overline{\sigma} \leq \overline{\tau}$, i.e., for all $\mathcal{S}$, for all P: if $\mathcal{S} \vDash \overline{\sigma} P$, then $\mathcal{S} \vDash \overline{\tau} P$. Choose a plausibility order $\mathcal{S}$ and a proposition P. By idempotence of attitudes, $\mathcal{S}^{\sigma P} = (\mathcal{S}^{\sigma P})^{\sigma P}$. So $\mathcal{S}^{\sigma P} \vDash \overline{\sigma} P$. By the initial assumption, $\mathcal{S}^{\sigma P} \vDash \overline{\tau} P$, so $(\mathcal{S}^{\sigma P})^{\tau P} = \mathcal{S}^{\sigma P}$. We have thus established that $\sigma P \cdot \tau P = \sigma P$, which concludes the right to left direction, and the proof. ⊣

So subsumption expresses inclusion of the corresponding fixed points and thus *inclusion of uninformativeness*. That $\sigma \leq \tau$ means that whenever applying σ is redundant, then so is applying τ, i.e., whenever a plausibility order is a fixed point of the upgrade σP, then it is also a fixed point of the upgrade τP.

1.8.5. PROPERTIES OF THE SUBSUMPTION ORDER. We observe a number of basic properties of our two strength orders. Let us first fix some standard terminology. As usual, a *preorder* is a reflexive and transitive binary relation; a *partial order* is an antisymmetric preorder.[22] Now let $\mathcal{O} = (O, \leq)$ be a preorder, and let $K \subseteq O$. Then $x \in O$ is a *lower bound* for K iff for all $y \in K$: $x \leq y$; and x is a *greatest lower bound* for K iff x is a lower bound for K and for all lower bounds y for K: $y \leq x$. Similarly, x is an *upper bound* for K iff for all $y \in K$: $y \leq x$; and x is a *least upper bound* for K in $\mathcal{O}$ iff x is an upper bound for K and for all upper bounds y for K: $x \leq y$. Observe that greatest lower bounds and least upper bounds are *unique* if $\mathcal{O}$ is a partial order, but not necessarily so if $\mathcal{O}$ is a preorder. Finally, a *lattice* is a partial order $\mathcal{O}$ such that for any $x, y \in O$: $\{x, y\}$ has a least upper bound and a greatest lower bound.

Using this terminology, the entailment and the subsumption order introduced above are distinguished primarily by the fact that the latter fails to be antisymmetric:

PROPOSITION 12.

1. *The entailment order on introspective propositional attitudes is a lattice.*

2. *The subsumption order on dynamic attitudes is a preorder $\mathcal{O} = (O, \leq)$ such that for any $x, y \in O$: $\{x, y\}$ has a (not necessarily unique) least upper bound and a (not necessarily unique) greatest lower bound.*

[22] A binary relation R is antisymmetric iff for any x and y in the domain of R: if xRy and yRx, then $x = y$.

3. *The subsumption order on dynamic attitudes is not a partial order, hence not a lattice.*

PROOF.

1. The entailment order on introspective propositional attitudes is obviously a partial order. Furthermore, given two attitudes A and A', the (unique) least upper bound for A and A' is given by $A \wedge A'$, while the (unique) greatest lower bound for A and A' is given by $A \vee A'$. So the entailment order is a lattice.

2. The subsumption order on dynamic attitudes is obviously a preorder. Now let τ and τ' be dynamic attitudes. Observe that the fixed point of the conjunction test $?(\tau \wedge \tau')$ is $\overline{\tau} \wedge \overline{\tau}'$. But $\overline{\tau} \wedge \overline{\tau}'$ is the (unique) least upper bound for $\overline{\tau}$ and $\overline{\tau}'$. It follows using Proposition 11 that $?(\tau \wedge \tau')$ is a (not necessarily unique) least upper bound for τ and τ'. For greatest lower bounds, we argue analogously using the disjunction test $?(\tau \vee \tau')$. Our claim follows.

3. We exhibit a counterexample, showing that the subsumption order is not antisymmetric. The fixed point of $?K$ is the same as the fixed point of $!$, i.e., irrevocable knowledge K. By Proposition 11, $?K \approx !$. However, obviously not $?K = !$. So the subsumption order is not antisymmetric, and thus not a partial order, hence not a lattice either. ⊣

The main point of the previous result is that the map sending dynamic attitudes to their fixed points is not one-to-one; given a propositional attitude A, it is not usually possible to identify a *unique* attitude τ such that $\overline{\tau} = A$.[23] In other words: there are *more* dynamic attitudes than introspective propositional ones, and as a result the subsumption order fails to be a lattice by Proposition 11.

1.8.6. QUESTION. If we understand introspective propositional attitudes as *targets for belief change* (cf. the previous section), the preceding observation raises an important question. Having chosen a suitable such target, how would an agent choose an appropriate dynamic attitude that realizes it? Is the choice completely arbitrary? Or, to put it differently: given distinct dynamic attitudes τ and τ' such that $\overline{\tau} = \overline{\tau}' = A$, can a case be made that one of them is, in some important respect, a *better* choice for realizing A? We will study one answer to this question in Chapter 3, where we use the criterion of *minimal*

[23] There are exceptions: the only dynamic attitude whose fixed point is absurdity $\bot$ is isolation $\varnothing$.

change that has traditionally been important in belief revision theory to select attitudes that are "more optimal" for a given fixed point over ones that are less so.

Chapter 2.
Trust, Distrust, Semi-Trust

The purpose of this chapter is to identify various classes of dynamic attitudes that formalize realiability assessments that are *trusting*, *distrusting*, or intermediate between these two extremes.

Our starting point is the notion of acceptance. To say that someone accepts that P, for some given proposition P, may have a static and a dynamic meaning, referencing either a state in which P is accepted, or an event of coming to accept P.[1] I will assume, following Stalnaker (1984), that to accept a proposition (i.e., the static sense) means "to treat the proposition as a true proposition" (ibid.). I deviate from Stalnaker's use of the term, however, in that I will assume that an agent who accepts that P is, at least for the moment, *committed* to P, even though this commitment may be defeasible.[2] So we can say that accepting that P amounts to being committed to treat P as a true proposition. This suggests a dynamic reading as well: on the dynamic reading, accepting a proposition P means "to come to be committed to treat the proposition P as a true proposition," or, as I will say for short, "to come to be committed to P."

Returning to the concept of acceptance in the static sense, the question is: what does "being committed to treat a proposition as a true proposition" amount to on a formal level? One approach would be to say that for such a commitment to be in place one needs to have hard information that P. On this view, an agent would be committed to P iff she has been able to exclude all

[1] The observation that there is a "pervasive ambiguity in our language between *products* and *activities*" (van Benthem 2011, emphasis in the original) was one of the starting points of van Benthem's program of "exploring logical dynamics" (van Benthem 1996). In our context, a state of acceptance is the product, and progressing to such a state through an event of accepting is the activity.

[2] Stalnaker (ibid.), on the other hand, subsumes propositional attitudes like "assuming" and "supposing" under the heading of acceptance. These attitudes do not necessarily involve any commitment on behalf of the agent, so the concept of acceptance I will use here is genuinely different.

non-P-worlds from her epistemic state, as represented by her plausibility order. However, this seems overly strong, as it ignores the fact that, in practice, it is often soft information we need to rely on: information that is defeasible but may still guide our actions. Taking soft information into account, the weakest possible formal interpretation we can give to the notion of acceptance—in the static sense—corresponds to our notion of *simple belief*: an agent accepts that P, in Stalnaker's sense of treating P as a true proposition, iff all the best worlds in her current plausibility order are P-worlds. This, then, will be our baseline notion of static acceptance: an agent accepts P in a plausibility order $\mathcal{S}$ iff $\mathrm{best}\,\mathcal{S} \subseteq P$; in other words: we cash in the informal notion of accepting a proposition, spelled out as being committed to treat that proposition as true, using the formal notion of simple belief.

In this chapter, I will take this static conception of acceptance for granted, and focus on its dynamic counterpart: what does it mean to *come to accept* information received from a source? Having a notion of dynamic acceptance at our disposal is useful for our purposes, since it allows us to develop a notion of epistemic trust according to which trusting a source means accepting the information received from that source.

We begin the work of this chapter by putting a notion of *uniform* trust in place: this is the topic of §2.1 and §2.2. We then consider variations on our theme: in §2.3, we discuss which dynamic attitudes formalize a notion of uniform distrust, and in §2.4, we discuss "semi-trusting" dynamic attitudes, which do not induce belief, but lead the recipient to suspend disbelief in the information received. On the basis of this work, §2.5 identifies seven qualitative degrees of trust and semi-trust.

§2.6 takes up a topic mentioned in the introduction of this dissertation. Realistically, sources are not trusted uniformly, but depending on the context, in particular, on the content of the information received.[3] In §2.6, we consider an operation on dynamic attitudes (called "mixture") that allows us to derive "mixed" forms of trust from the "uniform" ones we have considered in earlier sections of this chapter.

§§2.7–2.8 broaden the focus in another direction, considering a multi-agent extension of the setting building on the work of the earlier sections. This extension brings into view properties of communication acts made by an agent ("the speaker") in the presence of other agents ("the hearers") that assess the reliability of the speaker in potentially different ways. By means of a number of examples, we discuss how the setting models epistemic norms of communication, i.e., normative standards to which speakers can be held.

[3] Recall the example from the introduction to this dissertation: a mathematician may, for example, be trusted on mathematical, but not on administrative matters.

2.1. Positive Attitudes

Recall our example: if we receive the information that there are tigers in the Amazon jungle from a trusted source, we *accept* this information, in a *dynamic* sense: we transform our epistemic state in a certain way; the outcome of this transformation is a *state* of acceptance, a state in which we accept that there are tigers in the Amazon jungle, or, in Stalnaker's formulation: we "treat it as a true proposition" (Stalnaker 1984) that there are tigers in the Amazon jungle.

2.1.1. Coming to Accept. How are we to make sense of the idea of "coming to accept" something? We can take the notion of static acceptance outlined in the introduction to this chapter as a guideline: given our assumption that what is statically accepted by an agent is captured by the most plausible worlds in her plausibility order, the result of coming to accept that there are tigers in the Amazon jungle should be a epistemic state (represented by a plausibility order) in which all the most plausible worlds are worlds in which there are tigers in the Amazon jungle.

According to this answer, an agent whose original epistemic state is given by the order $\mathcal{S}$, and who (dynamically) accepts an input P from a τ-trusted source, will transform her epistemic state to an order $\mathcal{S}^{\tau P}$ such that

$$\text{best}\,\mathcal{S}^{\tau P} \subseteq P.$$

If this holds true in general of a dynamic attitude τ, then we say that τ satisfies *success*, i.e., τ satisfies success iff for any plausibility order $\mathcal{S}$ and proposition P: $\text{best}\,\mathcal{S}^{\tau P} \subseteq P$.

On the other hand, as much as our agent may trust the source, she has also another vital interest: she wants to maintain a consistent epistemic state. Thus, regardless of how much a particular source is trusted, our agent is interested in transforming her epistemic state in such a way so as to make sure that, under the condition that her epistemic state is consistent, it remains consistent, i.e.,

$$\text{if } \mathcal{S} \neq \varnothing \text{ then } \mathcal{S}^{\tau P} \neq \varnothing.$$

If this holds true in general of a dynamic attitude τ, then we say that τ satisfies *sanity*, i.e., τ satisfies sanity iff for any plausibility order $\mathcal{S}$ and proposition P: if $\mathcal{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$.

If we want to ensure *both* sanity and success in general, however, we run into a problem:

Proposition 13. *There exists no dynamic attitude τ satisfying success and sanity.*

PROOF. Towards a contradiction, let τ be a dynamic attitude satisfying success and sanity. Let $\mathcal{S}$ be a non-empty plausibility order, and let P be a proposition such that $P \cap \mathrm{S} = \varnothing$. By success, $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$. Since also $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq \mathrm{S}$, it follows that $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P \cap \mathrm{S}$, thus $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq \varnothing$, hence $\mathcal{S}^{\tau P} = \varnothing$. However, since $\mathcal{S} \neq \varnothing$, it follows from sanity that $\mathcal{S}^{\tau P} \neq \varnothing$. We have arrived at a contradiction, hence τ does not satisfy both sanity and success. ⊣

Given this tension between success ("coming to believe across the board") and sanity ("maintaining consistency across the board"), we cannot define acceptance by simply conjoining the two criteria.

There are a number of ways out. Most obviously, we could weaken one of the two criteria. So we could require that, for any order $\mathcal{S}$ and proposition P:

$$\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P \text{ and if } P \cap \mathrm{S} \neq \varnothing, \text{ then } \mathcal{S}^{\tau P} \neq \varnothing.$$

This means to demand success even at the cost of sanity: according to the requirement, an agent is to maintain consistency of her epistemic state only when receiving an input that is consistent with her hard information. This is the route taken in AGM theory—more in §2.2. On the other hand, one might just as well go into the other direction, and require:

$$\text{if } \mathcal{S} \neq \varnothing, \text{ then } \mathcal{S}^{\tau P} \neq \varnothing \text{ and if } P \cap \mathrm{S} \neq \varnothing, \text{ then } \mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P.$$

Working with this requirement, one sacrifices success, rather than sanity, in case the two are in conflict: we require generally that upgrades τP map non-empty plausibility orders $\mathcal{S}$ to non-empty plausibility orders $\mathcal{S}^{\tau P}$; and we require that the most plausible worlds after an upgrade τP are a subset of P given that there are P-worlds in $\mathcal{S}$.

Since there seems to be no compelling reason to favour one solution over the other, the formalization of acceptance we introduce below allows our agent to use *both* ways of resolving a possible conflict, the salomonic solution, as it were.

2.1.2. POSITIVE ATTITUDES. An attitude τ is *positive* if the following holds, for any given order $\mathcal{S}$ and proposition P:

$$\text{If } P \cap \mathrm{S} \neq \varnothing, \text{ then } \mathrm{S}^{\tau P} \neq \varnothing \text{ and } \mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P.$$

Examples of positive attitudes we have already seen are infallible trust !, strong trust $\Uparrow$ and minimal trust $\uparrow$.

According to the general definition, for any positive attitude τ, the upgrade τP will lead to a consistent epistemic state (i.e., a non-empty plausibility order) in which P is believed, as long as P is consistent with the hard

information of the agent before the upgrade. So, to reiterate, positive attitudes τ capture a specific form of *uniform trust*: the agent reacts to the proposition P received from a positively trusted source by *accepting* P, i.e., by *coming to believe that P—if P* is consistent with the agent's hard information. Moreover, positivity captures a form of *rational* acceptance: at least as long as this is possible, the agent will maintain a consistent epistemic state.

We ought to check if our definition of positive dynamic attitudes is in the spirit of the "salomonic solution" discussed above. This turns indeed out to be the case, given the background constraints in place in our framework. The following observation uses the *strong informativity* requirement:

PROPOSITION 14. *Let τ be positive. For any order $\mathcal{S}$ and proposition P, one of the following holds:*

1. $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$ *and if* $P \cap \mathrm{S} \neq \varnothing$, *then* $\mathcal{S}^{\tau P} \neq \varnothing$.

2. *if* $\mathcal{S} \neq \varnothing$, *then* $\mathcal{S}^{\tau P} \neq \varnothing$ *and if* $P \cap \mathrm{S} \neq \varnothing$, *then* $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$.

PROOF. Let τ be a positive dynamic attitude, let $\mathcal{S}$ be a plausibility order, and let P be a proposition. Suppose first that $P \cap \mathrm{S} \neq \varnothing$. By definition of positive attitudes, $\mathcal{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$. Hence both requirement (1.) and requirement (2.) are satisfied. Now suppose that $P \cap \mathrm{S} = \varnothing$. By strong informativity, either $\mathcal{S}^{\tau P} = \mathcal{S}$, in which case the requirement in (2.) is satisfied, or $\mathcal{S}^{\tau P} = \varnothing$, in which case the requirement in (1.) is satisfied. ⊣

So if the information received is *inconsistent* with what the agent already knows, "having a positive attitude" may mean either of two things: (1) the agent acquires inconsistent beliefs (i.e., she favours "success" over "sanity"); (2) the agent ignores the informational input (maintaining "sanity", but sacrificing success).

One also wonders what happens if the information received is *trivial* against the agent's current body of knowledge, i.e., what happens if the agent receives the information that P from a positively trusted source, and the agent's plausibility order $\mathcal{S}$ is such that $\mathrm{S} \subseteq P$? In that case, applying a positive attitude will leave the current order unaffected. This follows again from strong informativity:

PROPOSITION 15. *For any positive attitude τ:*

$$\textit{If } P \cap \mathrm{S} = \mathrm{S}, \textit{ then } \mathcal{S}^{\tau P} = \mathcal{S}.$$

PROOF. Let τ be positive and suppose that $P \cap \mathrm{S} = \mathrm{S}$. If $\mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing = \mathcal{S}$, and the claim holds. If $\mathrm{S} \neq \varnothing$, it follows by definition of τ that $\mathcal{S}^{\tau P} \neq \varnothing$. By strong informativity, $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$; so $\mathcal{S}^{\tau P} = \mathcal{S}$. ⊣

One may expect that positive attitudes create belief. However, given the way we have resolved the tension between "success" and "sanity", this is not actually the case. What we can say, however, is that any positive attitude creates the disjunction of opposite knowledge and belief:

PROPOSITION 16. *Positive attitudes τ create the disjunction of belief and opposite knowledge, i.e., if τ is positive, then for any order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \models BP \vee K^{\neg}P$.*

PROOF. From left to right, let τ be a positive attitude. If $P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \mathcal{S}$ or $\mathcal{S}^{\tau P} = \varnothing$ by the informativity of dynamic attitudes. In either case, $\mathcal{S}^{\tau P} \models K^{\neg}P$ and the claim holds. Suppose now that $P \cap \mathrm{S} \neq \varnothing$. By definition of τ, this implies that $\mathrm{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$. Hence $\mathcal{S}^{\tau P} \models BP$ and, again, the claim holds. ⊣

2.2. *Strictly Positive Attitudes*

Having a basic notion of (dynamic) acceptance in place—given by the positive dynamic attitudes, we now begin to consider variations on our theme.

2.2.1. STRICTLY POSITIVE ATTITUDES. A dynamic attitude τ is *strictly positive* iff the following are satisfied:

1. $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$.

2. If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$.

Adopting a strictly positive attitude towards a source means favouring "success" over "sanity" (cf. the discussion in the previous section). Since, as one can easily check, strictly positive attitudes are positive, the former represent, indeed, a *strict* form of acceptance within the latter wider class. If the input received from a strictly positively trusted source is inconsistent with what the agent already knows, then the agent acquires inconsistent beliefs:

PROPOSITION 17. *For any strictly positive attitude τ:*

$$\textit{If } P \cap \mathrm{S} = \varnothing, \textit{ then } \mathcal{S}^{\tau P} = \varnothing.$$

We take it that an upgrade to an inconsistent epistemic state will not happen "in practice", in the sense that an agent will avoid to perform such an upgrade, or, put differently, the upgrade is not actually *executable*, both intuitively and in the sense of our formal definition of executability (cf. §1.3). So

a different view on the matter is that information received from a strictly positively trusted source is *always* consistent with what the agent already knows. Holding a strictly positive attitude towards a source can thus be seen as *limiting the range of information that can originate from that source* by virtue of the fact that upgrades given by propositions that are inconsistent with the hard information of the agent are unexcutable.

As a consequence of the previous proposition, strictly positive attitudes create belief:

PROPOSITION 18. *Strictly positive attitudes τ create belief, i.e., if τ is strictly positive, then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \models BP$.*

PROOF. Let τ be a strictly positive attitude. Suppose that $\mathrm{S} \cap P = \varnothing$. By the previous proposition, $\mathcal{S}^{\tau P} = \varnothing$, so $\mathcal{S}^{\tau P} \models BP$ and the claim holds. On the assumption that $\mathrm{S} \cap P \neq \varnothing$, we argue as in Proposition 16 to conclude that $\mathcal{S}^{\tau P} \models BP$ and, again, the claim holds. ⊣

2.2.2. AGM OPERATORS. Let us clarify how the notion of (strict) positivity relates to the traditional AGM postulates.[4] The AGM postulates were originally stated in a purely syntactic framework. In translating them into our semantic setting, we follow the formulation of Robert Stalnaker.[5]

Let $\star$ be a dynamic attitude. Then $\star$ is an *AGM (revision) operator* iff $\star$ satisfies, for any plausibility order $\mathcal{S}$ and proposition P:

1. $\mathrm{best}\,\mathcal{S}^{\star P} \subseteq P$. (*success*)

2. If $\mathrm{best}\,\mathcal{S} \cap P \neq \varnothing$, then $\mathrm{best}\,\mathcal{S}^{\star P} = \mathrm{best}\,\mathcal{S} \cap P$. (*expansion*)

3. If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\star P} \neq \varnothing$. (*conditional sanity*)

4. If $\mathrm{best}\,\mathcal{S}^{\star P} \cap Q \neq \varnothing$, then $\mathrm{best}\,\mathcal{S}^{\star (P \cap Q)} = \mathrm{best}\,\mathcal{S}^{\star P} \cap Q$. (*rational monotonicity*)

So the AGM operators form a subclass of the strictly positive attitudes. Since a dynamic attitude τ is strictly positive iff it satisfies the first and third

[4] Cf. Alchourrón et al. (1985), Gärdenfors (1988).

[5] Cf. Stalnaker (2009). Stalnaker's footnote 5 clarifies how the semantic formulation precisely relates to the original postulates: the postulates, as listed here, are equivalent regroupings of the original AGM postulates. Furthermore, two of the original AGM postulates are missing: the one which says that logically equivalent sentences induces the same belief changes; and the one according to which the output of performing a revision should be a deductively closed set of sentences (a "theory"). These two postulates are unnecessary in a purely semantic setting.

postulate in the above list—*success* and *conditional sanity*—, τ is an AGM operator iff it is strictly positive and, in addition, satisfies the second and fourth postulate—*expansion* and *rational monotonicity*.

The *positive* attitudes (simpliciter), on the other hand, arise from the above list by replacing the *success* postulate with the weaker requirement of *conditional success*: "If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{best}\, \mathcal{S}^{\tau P} \subseteq P$."

2.2.3. Examples. Infallible trust ! is an AGM operator; however, strong trust $\Uparrow$ and minimal trust $\uparrow$ are not, for the simple reason that they are not strictly positive, violating *success of revision*. In the Belief Revision literature studying operations on Grove spheres (i.e., plausibility orders), the fact that the most well-known of these operations (for example, Boutilier (1993)'s minimal revision, i.e., our $\uparrow$, and Nayak (1994)'s lexicographic revision, our $\Uparrow$) do not satisfy the success postulate is not usually stressed. The reason is, perhaps, that it is not difficult to obtain strictly positive, and indeed, AGM "versions" of both strong and minimal trust (i.e., the attitudes underlying lexicographic and minimal revision, respectively). To this end, we introduce an operation converting positive into strictly positive attitudes, which we call "stricture".

2.2.4. Strictures. Let τ be a positive dynamic attitude. The *stricture* τ^+ of τ is defined by means of

$$\mathcal{S}^{\tau^+ P} := \begin{cases} \mathcal{S}^{\tau P} & P \cap \mathrm{S} \neq \varnothing, \\ \varnothing & \text{otherwise.} \end{cases}$$

The stricture operator *limits* the information that can be received from a source to propositions P that are not already known to be false by the recipient. If, on the other hand, P *is* already known to be false, the agent comes to accept P anyway, but at the expense of ending up in an inconsistent epistemic state.

Observe that for strictly positive τ, we have, by Proposition 17, that $\tau^+ = \tau$. More generally, the strictly positive attitudes can be characterized as the strictures of positive attitudes:

proposition 19. *An attitude τ is strictly positive iff τ is the stricture of some positive attitude.*

proof. Suppose that τ is strictly positive. Then, as just observed, $\tau^+ = \tau$, and since τ is positive, τ is the stricture of some positive attitude, namely the stricture of τ. Conversely, suppose that τ is the stricture of some positive attitude, say σ. We need to show that τ is strictly positive. So let $\mathcal{S}$ be a plausibility order, and let P be a proposition. We have to show that (1) $\mathrm{best}\, \mathcal{S}^{\tau P} \subseteq P$, and

that (2) if $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$. For (1), observe that if $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{\sigma P}$, and since σ is positive, the claim holds. Further, if $P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing$ by definition of strictures, and again, the claim holds. For (2), observe that if $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{\sigma P}$, and since σ is positive, the claim holds. So τ is strictly positive. ⊣

In particular, the stricture of minimal trust, $\uparrow^+$, is called *strict minimal trust*, and the stricture of strong trust, $\Uparrow^+$, is called *strict strong trust*.

PROPOSITION 20. *$\uparrow^+$ and $\Uparrow^+$ are AGM operators.*

PROOF. We check the claim for the case of $\uparrow^+$. Let $\mathcal{S}$ be a plausibility order, and let P be a proposition.

1. If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{best}\, \mathcal{S}^{\uparrow P} \subseteq P$ by definition of $\uparrow$, and since, in that case, $\mathcal{S}^{\uparrow^+ P} = \mathcal{S}^{\uparrow P}$, also $\mathrm{best}\, \mathcal{S}^{\uparrow^+ P} \subseteq P$. If, on the other hand, $P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\uparrow^+ P} = \varnothing$ by definition of $\uparrow^+$, so $\mathrm{best}\, \mathcal{S}^{\uparrow^+ P} \subseteq P$. It follows that $\uparrow^+$ satisfies success.

2. If $\mathrm{best}\, \mathcal{S} \cap P \neq \varnothing$, then $\mathrm{best}\, \mathcal{S}^{\uparrow^+ P} = \mathrm{best}\, \mathcal{S} \cap P$ by definition of $\uparrow^+$. So $\uparrow^+$ satisfies expansion.

3. If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{best}\, \mathcal{S}^{\uparrow^+ P} \neq \varnothing$, since in that case, $\mathrm{S}^{\uparrow^+ P} = \mathrm{S}$. So $\uparrow^+$ satisfies conditional sanity.

4. Finally, suppose that $\mathrm{best}\, \mathcal{S}^{\uparrow^+ P} \cap Q \neq \varnothing$. Thus, by definition of $\uparrow^+$, $(\mathrm{best}_{\mathcal{S}}\, P) \cap Q \neq \varnothing$. Hence $\mathrm{best}_{\mathcal{S}}(P \cap Q) = (\mathrm{best}_{\mathcal{S}}\, P) \cap Q$. By definition of $\uparrow^+$, $\mathrm{best}\, \mathcal{S}^{\uparrow^+ (P \cap Q)} = \mathrm{best}_{\mathcal{S}}(P \cap Q)$, but as we have just seen, $\mathrm{best}_{\mathcal{S}}(P \cap Q) = (\mathrm{best}_{\mathcal{S}}\, P) \cap Q$, so $\mathrm{best}\, \mathcal{S}^{\uparrow^+ (P \cap Q)} = \mathrm{best}\, \mathcal{S}^{\uparrow^+ P} \cap Q$. Thus $\uparrow^+$ satisfies rational monotonicity. ⊣

By Proposition 18, $\uparrow^+$ creates belief; it is also easy to see that $\Uparrow^+$ creates *strong belief*. The stricture operation thus provides a remedy for the disharmony between $\uparrow$ and B on the one hand, and $\Uparrow$ and Sb on the other hand that we observed earlier (cf. §1.4.4 and in particular Proposition 4). Going the other direction, one would like to have a "version" of infallible trust that is positive, but not strictly positive. For this purpose, we define the attitude ¡, called *weak infallible trust*, by means of

$$\mathcal{S}^{¡P} := \begin{cases} \mathcal{S}^{!P} & P \cap \mathrm{S} \neq \varnothing, \\ \mathcal{S} & P \cap \mathrm{S} = \varnothing. \end{cases}$$

This attitude has the desired properties: it is positive, but not strictly positive (and thus not an AGM operator); moreover, its stricture is infallible trust !. Information received from a weakly infallibly trusted source comes with a

qualified warranty of truthfulness: the agent acquires hard information in the proposition received, unless she already (infallibly) knows that that proposition is not satisfied; in the latter case, she simply ignores the input.

The difference between positive attitudes and their strictures is slight and may be viewn as somewhat "technical"; it is nevertheless significant. We return to the issue in §2.5 below.

2.3. *Negative Attitudes*

There is a wide range of reliability assessments that are non-trusting. Most obviously, there is *distrust*. In Raymond Smullyan's famous logic puzzles, one even encounters sources of information that are best dealt with by *uniformly distrusting* them.

> A Smullyan liar claims: *I don't live on this island.*

As anyone familiar with Smullyan's logic puzzles knows,[6] information received from "Smullyan liars" comes with a *warranty of falsehood*: the liars in the puzzle are sources of information that are *predictably wrong*: whatever they say, the opposite is true. Having identified a liar as as a liar (which is, of course, the main difficulty involved in solving the puzzles), the best strategy is to upgrade one's beliefs to the contrary when receiving information from him: so if the liar says he does not live on this island, one should conclude that he does. This may still be seen as a form of acceptance: rather than accepting the proposition received, one accepts its complement—think of it as "negative acceptance".

2.3.1. Negative Attitudes. Formally, an attitude τ is *negative* if (for any $\mathcal{S}$ and P)

$$\text{if } S \cap \neg P \neq \varnothing, \text{ then } S^{\tau P} \neq \varnothing \text{ and } \operatorname{best} \mathcal{S}^{\tau P} \subseteq \neg P.$$

On the other hand, an attitude τ is *strictly negative* if the following holds:

1. If $S \cap \neg P \neq \varnothing$, then $S^{\tau P} \neq \varnothing$.
2. $\operatorname{best} \mathcal{S}^{\tau P} \subseteq \neg P$.

The distinction between negative and strictly negative attitude parallels the distinction between positive and strictly positive ones. We obtain analogues to earlier observations:

[6] Cf., for example, Smullyan (1978).

PROPOSITION 21.

1. *For any negative attitude τ: If $\neg P \cap \mathrm{S} = \mathrm{S}$, then $\mathcal{S}^{\tau P} = \mathcal{S}$.*

2. *For any strictly negative attitude τ: If $\neg P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing$.*

PROOF. The first item is analogous to Proposition 15; the second item is analogous to Proposition 17. ⊣

PROPOSITION 22.

1. *Negative attitudes τ create the disjunction of disbelief and knowledge, i.e., if τ is negative, then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \models B^{\neg} P \vee KP$.*

2. *Strictly negative attitudes τ create disbelief, i.e., if τ is strictly negative, then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \models B^{\neg} P$.*

PROOF. The first item is analogous to Proposition 16; the second item is analogous to Proposition 18. ⊣

Observe that the (strictly) negative attitudes are just the *opposites* of the (strictly) positive attitudes, where we recall from Chapter 1.5, that the opposite $\tau^{\neg}$ of an attitude τ is given by $\tau^{\neg} P := \tau(\neg P)$.

PROPOSITION 23. *An attitude τ is (strictly) positive iff its opposite $\tau^{\neg}$ is (strictly) negative.*

PROOF. We show the claim for strictly positive attitudes; the claim for positive attitudes is shown analogously. Suppose that τ is strictly positive, and consider a plausibility order $\mathcal{S}$ and a proposition P. Since τ is strictly positive, $\mathrm{best}\, \mathcal{S}^{\tau(\neg P)} \subseteq \neg P$. Furthermore, if $\neg P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau(\neg P)} \neq \varnothing$. Recalling that $\tau^{\neg} P = \tau(\neg P)$ for any P, it follows that $\tau^{\neg}$ is strictly negative. ⊣

2.3.2. EXAMPLES. Using the opposite operator, we can define various new dynamic attitudes in terms of our earlier examples. In particular, *infallible distrust* is given by $!^{\neg}$, *strong distrust* is given by $\Uparrow^{\neg}$, *strict strong distrust* is given by $\Uparrow^{-} := (\Uparrow^{+})^{\neg}$, *minimal distrust* is given by $\uparrow^{\neg}$, and *strict minimal distrust* is given by $\uparrow^{-} := (\uparrow^{+})^{\neg}$. Of these, $\Uparrow^{\neg}$ and $\uparrow^{\neg}$ are negative, while $!^{\neg}$, $\Uparrow^{-}$ and $\uparrow^{-}$ are strictly negative.

2.4. Semi-Positive Attitudes

Quite common in daily life are occasions where a source is—what we shall call—"semi-trusted". Consider an example:

> The weather forecast: *There will be heavy rainfall tomorrow.*

Many people treat the weather forecast as a *source of positive but inconclusive evidence*, in the following sense: if their prior beliefs conflict with what the forecast says, rather than coming to believe in the forecast, they become *unopinionated* on the relevant bit of information. Suppose you believe that the weather will be fine tomorrow. Upon hearing the forecast predict heavy rainfall, you would, assuming a semi-trusting attitude, *lose your belief* that the weather will be fine tomorrow, but *without coming to believe* that there will be heavy rainfall either. You would consider both possibilities as plausible: fine weather, and heavy rainfall. This may influence your decisions: you might decide to take an umbrella, *just in case*. While this should certainly not be seen as "fully accepting" the information received, the latter is not rejected either. Think of the phenomenon just described as "partial acceptance", or "semi-acceptance" of the information received.

Hearers often entertain attitudes of this kind towards sources that are assumed to be cooperative, but whose competence on a specific issue may still raise some doubts. Even if a helpful stranger is doing his best to indicate the correct way to the station, she might not precisely know it herself. Then, our best bet might be to take her directions as one plausible option (while keeping open the possibility that another route might be the correct one). Similarly, even if the weather forecast is trying to inform us correctly, the fact that they have been wrong before makes "semi-trusting" them seem like a useful strategy to adopt.

2.4.1. Semi-Positive Attitudes. An attitude τ is *semi-positive* iff:

$$\text{If } P \cap \mathrm{S} \neq \varnothing, \text{ then } \operatorname{best} \mathcal{S}^{\tau P} \cap P \neq \varnothing.$$

Observe that positive attitudes are semi-positive: the class of semi-positive attitudes is *wider*. Intuitively, an attitude is semi-positive if its propositional argument P is *at least taken into account* by the agent, as far as its plausibility goes. P will register "semi-positively" in the agent's epistemic state in the sense that the set of plausible worlds is guaranteed to contain some P-worlds after the upgrade, unless P was already known to be false: this is exactly what is captured by the above formal requirement that if $P \cap \mathrm{S}$ is non-empty, then

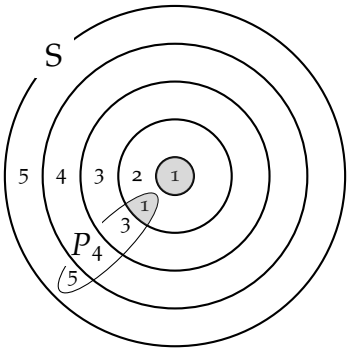


FIGURE 6. Semi-trust $\uparrow^{\sim}$ applied to P: the upgrade $\uparrow^{\sim}P$ adds the most plausible P-worlds in the original order to the most plausible worlds overall in the original order.

also $\text{best}_{\mathcal{S}^{\tau P}} \cap P$ is non-empty. So the source is taken to provide genuine, but inconclusive evidence (as in the case of the weather forecast).

In a situation where an agent believes but does not know that $\neg P$, receiving the information that P from a semi-positively trusted τ-source will result in *suspension of disbelief*, i.e., applying the upgrade τP will result in an order satisfying $B^{\sim}P$, the dual of belief in P (cf. §1.2.10) given by

$$\mathcal{S} \vDash BP \text{ iff } \text{best}\,\mathcal{S} \cap P \neq \varnothing.$$

In this way, semi-positive attitudes may, for example, be used to formalize our weather forecast example: upon receiving the information that it will rain tomorrow from the weather forecast, the agent merely comes to believe *that it may rain tomorrow*, i.e., after the upgrade, some of the most plausible worlds come to be worlds where it rains tomorrow.

2.4.2. EXAMPLE. For any proposition P, the *dual minimal upgrade* $\uparrow^{\sim}P$ adds the best P-worlds to the best worlds overall, leaving everything else unchanged; *semi-trust* $\uparrow^{\sim}$ is the dynamic attitude given by the family of upgrades $\{\uparrow^{\sim}P\}_{P\subseteq W}$. Figure 6 illustrates the behaviour of this dynamic attitude, which is a typical example of a semi-positive attitude.

2.4.3. APPLYING OPPOSITES AND STRICTURES. We define the *semi-negative* attitudes as the opposites of semi-positive attitudes; and the *strict semi-positive* (*strict semi-negative*) attitudes as the strictures of semi-positive (semi-negative) attitudes. In terms of semantic clauses, strictures and opposites of semi-positive attitudes may be described as follows:

PROPOSITION 24.

1. *An attitude τ is* semi-negative *iff: if $\neg P \cap \mathrm{S} \neq \varnothing$, then* $\mathrm{best}\, \mathcal{S}^{\tau P} \cap \neg P \neq \varnothing$.

2. *An attitude τ is* strictly semi-positive *iff: (1) if $P \cap \mathrm{S} \neq \varnothing$, then and* $\mathrm{best}\, \mathcal{S}^{\tau P} \cap P \neq \varnothing$, *and (2) if $P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing$.*

3. *An attitude τ is* strictly semi-negative *iff: (1) if $\neg P \cap \mathrm{S} \neq \varnothing$, then* $\mathrm{best}\, \mathcal{S}^{\tau P} \cap \neg P \neq \varnothing$, *and (2) if $\neg P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing$.*

PROOF. We prove the first item as an example. Let $\mathcal{S}$ be a plausibility order, and P a proposition. Suppose that $\neg P \cap \mathrm{S} \neq \varnothing$. By definition of τ, $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau^{\neg} \neg P}$, with $\tau^{\neg}$ semi-positive. Since $\neg P \cap \mathrm{S} \neq \varnothing$ and $\tau^{\neg}$ is semi-positive, we conclude that $\mathrm{best}\, \mathcal{S}^{\tau^{\neg} \neg P} \cap \neg P \neq \varnothing$. But since $\mathcal{S}^{\tau^{\neg} \neg P} = \mathcal{S}^{\tau P}$, we have shown our claim. ⊣

We also observe that semi-positive attitudes create the disjunction of opposite knowledge $K^{\neg}$ and the dual of belief $B^{\sim}$, while strictly semi-positive attitudes create the dual of belief $B^{\sim}$:

PROPOSITION 25.

1. *Semi-positive attitudes τ create the disjunction of opposite knowledge and dual belief, i.e., if τ is semi-positive, then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \vDash K^{\neg} P \vee B^{\sim} P$.*

2. *Positive attitudes τ create dual belief, i.e., if τ is semi-positive, then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \vDash B^{\sim} P$.*

PROOF. The first item is similar to Proposition 16; the second item is similar to Proposition 18. ⊣

2.4.4. AMBIVALENCE. As a side remark, observe that an attitude may well be semi-positive *and* semi-negative. One could call such attitudes "ambivalent". In a situation where neither P nor its complement $\neg P$ are known to be true by the agent, and she receives the information that P from a source towards which she has an ambivalent attitude, the agent's plausibility order will satisfy both $B^{\sim} P$ and $B^{\sim} \neg P$ after performing the corresponding upgrade. An ambivalently trusted source will thus lead the agent to become *uncertain about* P in this stronger sense: receiving the information that P "makes P an issue."

2.4.5. AGM Contraction Operators. Receiving a piece of information P from a semi-negatively trusted source will lead the hearer to *lose belief in* P (unless P is already known to be true). In other words, receiving the information that P from such a source is apt to *cast doubt on* the truth of P. As we will now see, in the terminology of belief revision theory, this means that semi-negative attitudes perform an operation of contraction.

Let us make precise the relation to the postulates for contraction imposed in the AGM literature.

An attitude τ is an *AGM contraction operator* iff for any order $\mathcal{S}$ and proposition P:

— $\mathrm{best}\,\mathcal{S} \subseteq \mathrm{best}\,\mathcal{S}^{\tau P}$. (*inclusion*)
— If $\mathrm{best}\,\mathcal{S} \cap \neg P \neq \varnothing$, then $\mathrm{best}\,\mathcal{S}^{\tau P} = \mathrm{best}\,\mathcal{S}$. (*vacuity*)
— If $\mathrm{S} \cap \neg P \neq \varnothing$, then $\mathrm{best}\,\mathcal{S}^{\tau P} \cap \neg P \neq \varnothing$. (*success of contraction*)
— If $\mathrm{best}\,\mathcal{S} \subseteq P$, then $\mathrm{best}\,\mathcal{S}^{\tau P} \cup (\mathrm{best}\,\mathcal{S} \cap P) \subseteq \mathrm{best}\,\mathcal{S}$. (*recovery*)
— $\mathrm{best}\,\mathcal{S}^{\tau(P\cap Q)} \subseteq \mathrm{best}\,\mathcal{S}^{\tau P} \cup \mathrm{best}\,\mathcal{S}^{\tau Q}$. (*conjunctive overlap*)
— If $\mathrm{best}\,\mathcal{S}^{\tau(P\cap Q)} \cap \neg P \neq \varnothing$, then $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq \mathrm{best}\,\mathcal{S}^{\tau(P\cap Q)}$. (*conjunctive inclusion*)

By virtue of the *success of contraction* postulate, the AGM contraction operators form a subclass of the semi-negative attitudes. A typical example of an AGM contraction operator is semi-distrust $(\uparrow^{\sim})^{\neg}$ (which is simply the opposite of semi-trust).

2.4.6. Reachability. One may wonder of what use sources that are semi-positively trusted could ever be for an agent. Isn't it more useful to receive all one's information from sources that are trusted (in the sense given by positive, or even strictly positive dynamic attitudes)? As we show in the remainder of the current section, the answer is not as straightforward. In particular, as we will see, receiving information from positively trusted sources only may constrain the possible future evolutions of an agent's information state. Having semi-trusted sources at one's disposal, on the other hand, provides a remedy for this problem.

We start by putting the necessary terminology in place. We define the relation $\twoheadrightarrow$ on dynamic attitudes by means of

$$\mathcal{S} \twoheadrightarrow \mathcal{S}' \text{ iff } \mathrm{S}' \subseteq \mathrm{S}.$$

Equivalently, $\mathcal{S} \twoheadrightarrow \mathcal{S}'$ iff there exists an upgrade u such that $\mathcal{S}^u = \mathcal{S}'$. If $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, then we say that $\mathcal{S}'$ is *in the dynamic scope of* $\mathcal{S}$.

Essentially, $\mathcal{S}'$ is in the dynamic scope of $\mathcal{S}$ if $\mathcal{S}'$ is reachable from $\mathcal{S}$ "in theory." But is $\mathcal{S}'$ also reachable from $\mathcal{S}$ "in practice"? What we mean by

this is the following. Consider an agent who has a number of sources of information at his disposal. Let us assume that the dynamic attitues towards the sources are collected in the set Δ. An interesting question to consider is whether any order $\mathcal{S}'$ that is in the dynamic scope of some given order $\mathcal{S}$ ("reachable in theory") can actually be "realized" by means of processing pieces of information received from the agent's sources, i.e., by a sequence of upgrades $\tau_0 P_0, \ldots, \tau_n P_n$, where each τ_k is an element of Δ ("reachable in practice"). If this is the case in general, then the set Δ may be said (as defined formally below) to be "dynamically complete": the agent has "enough sources" to be able to reach any order in the dynamic scope of his current order.

Let us spell this out formally. Given a set of dynamic attitudes Δ, we write $\mathcal{S} \twoheadrightarrow_\Delta \mathcal{S}'$ iff there exists an attitude $\tau \in \Delta$ and a proposition $P \subseteq W$ such that $\mathcal{S}^{\tau P} = \mathcal{S}'$. Notice that whenever $\mathcal{S} \twoheadrightarrow_\Delta \mathcal{S}'$ for some Δ, then also $\mathcal{S} \twoheadrightarrow \mathcal{S}'$.

We say that a set of dynamic attitudes Δ is *(dynamically) complete* if for any plausibility orders $\mathcal{S}$ and $\mathcal{S}'$ such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$ there exists a sequence of plausibility orders $\mathcal{S}_0, \ldots, \mathcal{S}_n$ ($n \geq 0$) such that $\mathcal{S} \twoheadrightarrow_\Delta \mathcal{S}_0 \twoheadrightarrow_\Delta \cdots \twoheadrightarrow_\Delta \mathcal{S}_n \twoheadrightarrow_\Delta \mathcal{S}'$.

Intuitively, this boils down to the picture painted right above: suppose an agent is in the epistemic state given by $\mathcal{S}$, and suppose that the set Δ collects the agent's sources of information. Take any order $\mathcal{S}'$ such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$. The question is if the sources can inform the agent, by sending him a sequence of pieces of information, in such a way that he ends up in the epistemic state given by $\mathcal{S}'$. If this is the case for arbitrary pairs $\mathcal{S}$ and $\mathcal{S}'$ such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, then the set of dynamic attitudes Δ is called dynamically complete. Otherwise, it is dynamically incomplete.

A first observation to make is that sets of *monotonic* dynamic attitudes are never complete. Let Δ be a set of dynamic attitudes. Δ is *monotonic* iff for any order $\mathcal{S}$, proposition P and $\tau \in \Delta$: if $\mathcal{S}^{\tau P} \neq \mathcal{S}$, then it is not the case that $\mathcal{S}^{\tau P} \twoheadrightarrow_\Delta \mathcal{S}$.

In other words: changes given by monotonic sets of dynamic attitudes are *irreversible*: there is no way to fall back to an earlier position once the order has been transformed in a particular way. An obvious example of a monotonic set of dynamic attitudes is given by the singleton set containing only infallible trust !: having deleted a world from a plausibility order, it is gone for good.

THEOREM 26. *Let Δ be a set of dynamic attitudes. If Δ is monotonic, then Δ is incomplete.*

PROOF. We show the contrapositive: if Δ is complete, then Δ is not monotonic. Suppose that Δ is complete. Let $\mathcal{S}$, $\mathcal{S}'$ be plausibility orders such that $\mathcal{S} \neq \mathcal{S}'$ and $\mathrm{S} = \mathrm{S}'$. By the fact that Δ is complete, we have $\mathcal{S} \twoheadrightarrow_\Delta \mathcal{S}'$, so

there exists a sequence of upgrades $\tau_1 P_1 \dots \tau_n P_n$ of minimal length such that $(\cdots(\mathcal{S}^{\tau_1 P_1})\cdots)^{\tau_n P_n} = \mathcal{S}'$, with $\tau_k \in \Delta$, and $P_k \subseteq W$ for $1 \leq k \leq n$. Note that $n \geq 1$ since $\mathcal{S} \neq \mathcal{S}'$. Since our sequence is of minimal length, $(\cdots(\mathcal{S}^{\tau_1 P_1})\cdots)^{\tau_{n-1} P_{n-1}} \neq \mathcal{S}'$. Since Δ is complete, $\mathcal{S}' \twoheadrightarrow_\Delta (\cdots(\mathcal{S}^{\tau_1 P_1})\cdots)^{\tau_{n-1} P_{n-1}}$. So Δ is not monotonic. ⊣

Not unexpectedly, then, non-monotonicity is a necessary criterion for a set of dynamic attitudes to be complete. More interestingly, sets of *positive* dynamic attitudes are never complete.

LEMMA 27. *Let w, v be possible worlds. Consider the plausibility order*

$$\mathcal{S} = (\{w, v\}, \{(w, v), (w, w), (v, v)\}).$$

For any positive dynamic attitude τ and proposition P: if $\mathrm{S}^{\tau P} = \mathrm{S}$, then $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \mathcal{S}'\}$, where $\mathcal{S}' = (\{w, v\}, \{(v, w), (w, w), (v, v)\})$.

PROOF. We assume the notation from the statement of the lemma. Suppose first that $P \cap \mathrm{S} = \varnothing$. By informativity of dynamic attitudes, $\mathcal{S}^{\tau P} \in \{\varnothing, \mathcal{S}\}$. Since $\mathrm{S} \neq \varnothing$, it follows that $\mathcal{S}^{\tau P} = \mathcal{S}$, and our claim holds. Suppose, second, that $P \cap \mathrm{S} \neq \varnothing$. We distinguish two sub-cases. First, suppose that $\{w, v\} \subseteq P$. Then $\mathcal{S}^{\tau P} = \mathcal{S}$ by Proposition 15, and our claim holds. Second, suppose that $\{w, v\} \nsubseteq P$. This implies that either (1) $w \in P, v \notin P$ or (2) $w \notin P, v \in P$. Assuming (1), it follows that $(w, v) \in \mathcal{S}^{\tau P}$, $(v, w) \notin \mathcal{S}^{\tau P}$ since τ is positive. But then, $\mathcal{S}^{\tau P} = \mathcal{S}$. Assuming (2), it follows that $(v, w) \in \mathcal{S}^{\tau P}$, $(w, v) \notin \mathcal{S}^{\tau P}$, again, since τ is positive. But then $\mathcal{S}^{\tau P} = \mathcal{S}'$. So assuming either of (1) or (2), our claim holds. Since we have considered all cases, the proof is complete. ⊣

THEOREM 28. *Any set of positive dynamic attitudes is dynamically incomplete.*

PROOF. Let Δ be a set of positive attitudes. Consider the order

$$\mathcal{S} = (\{w, v\}, \{(w, v), (w, w), (v, v)\}).$$

We claim that for any $n \geq 1$, and sequence of upgrades $\tau_1 P_1 \dots \tau_n P_n$ such that $\tau_k \in \Delta$ and $P_k \subseteq W$ for $1 \leq k \leq n$: if the domain of $(\cdots(\mathcal{S}^{\tau_1 P_1})\cdots)^{\tau_n P_n}$ is S, then

$$(\cdots(\mathcal{S}^{\tau_1 P_1})\cdots)^{\tau_n P_n} \in \{\mathcal{S}, \mathcal{S}'\},$$

where $\mathcal{S}' = (\{w, v\}, \{(v, w), (w, w), (v, v)\})$. We show the claim by an easy induction on n, using the previous lemma. Now we observe that none of the orders in $\{\mathcal{S}, \mathcal{S}'\}$ equals the plausibility order

$$\mathcal{S}'' = (\{w, v\}, \{(w, v), (v, w), (w, w), (v, v)\}).$$

Thus it is not the case that $\mathcal{S} \twoheadrightarrow_\Delta \mathcal{S}''$. However, $\mathcal{S} \twoheadrightarrow \mathcal{S}''$. So Δ is incomplete. ⊣

The problem with positive attitudes highlighted by the previous result is that they make it difficult to retreat to a "flatter" doxastic position, where two worlds w and v such that w had been strictly more plausible than v now become *equiplausible*. This, however, is just what certain semi-positive dynamic attitudes allow an agent to do:

THEOREM 29. *There exist dynamically complete sets of semi-positive dynamic attitudes.*

PROOF. We show that $\Delta = \{!, \uparrow, \uparrow^{\sim}\}$ is dynamically complete. Since $\uparrow$, $\uparrow^{\sim}$ and $!$ are semi-positive, this entails the original claim. Let $\mathcal{S}, \mathcal{S}'$ be plausibility orders such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$. We give a sketch how to obtain $\mathcal{S}'$ from $\mathcal{S}$ using $!$, $\uparrow$ and $\uparrow^{\sim}$: first, upgrade $\mathcal{S}$ with $!S'$; second, make all worlds in S' equiplausible using a sequence of upgrades given by $\uparrow^{\sim}$ and appropriately chosen propositions (essentially, we consecutively add worlds to the set of best worlds overall, until *all* worlds are best, that is, all worlds are equiplausible); third, construct the desired order $\mathcal{S}'$ using a sequence of upgrades given by $\uparrow$ and appropriately chosen propositions (essentially, this is done layer by layer: first, we put the set of worlds to the top that are least plausible in $\mathcal{S}'$, then the set of worlds that are second-to-least plausible in $\mathcal{S}'$, and so on, and in this way we are guaranteed to finally arrive at $\mathcal{S}'$ itself). As a result, $\mathcal{S} \twoheadrightarrow_{\Delta} \mathcal{S}'$. Hence $\Delta = \{\uparrow, \uparrow^{\sim}, !\}$ is dynamically complete. ⊣

The construction of $\mathcal{S}'$ from $\mathcal{S}$ crucially involves the use of $\uparrow^{\sim}$ to "flatten" any strict inequalities among worlds. This is just what *positive* dynamic attitudes do not allow an agent to do.

It follows from the preceding observations that an agent will only be able to reach, in general, *any* epistemic state from her current epistemic state if she has sources at her disposal that are *semi-trusted* but not *positively trusted*. In this sense, being surrounded by trusted friends only is not necessary a good thing!

2.5. Qualitative Degrees of Trust and Semi-Trust

It is time to summarize the results so far, and expand on them from a more high-level perspective. So far in this chapter, we have seen a number of main classes of dynamic attitudes. They are summarized in Table 1. All of them are closely related to the propositional attitude *simple belief* B. In particular, strictly positive attitudes (which are just the strictures of positive attitudes) create belief B; strictly negative attitudes (which are just the opposites of strictures of positive attitudes) create disbelief $B^{\neg}$; and strict semi-positive

attitudes (which are just the strictures of semi-positive attitudes) create dual belief $B^{\sim}$.

An analogous style of analysis may now be applied to other propositional attitudes, in particular, to the two main other ones we have seen: irrecovable knowledge K, and strong belief Sb. Instead of going through all the definitions in detail, we summarize the relevant definitions, results and examples in Table 2 and Table 3. The two tables make use of two *three new examples* of dynamic attitudes, which we discuss immediately below.

Table 2 highlights the seven basic classes of dynamic attitudes that are the outcome of our analysis; they correspond to seven distinct ways of making sense of the notion of acceptance in a qualitative sense. Put differently, they correspond to seven different forms of trust, where trust is conceptualized, as throughout this dissertation, as an assessment of reliability, encoded in our notion of a dynamic attitude. In view of the fact that irrevocable knowledge K, strong belief Sb, refined belief Rb and simple belief B seem to be the most natural propositional attitudes our setting gives rise to, we regard these as the fundamental forms of trust the setting can capture.

Notice that we are pushing the notion of acceptance to the limit: for a barely semi-positive attitude τ (the last row in Table 2), the propositional input P is "accepted" only insofar as applying the upgrade τP to an order $\mathcal{S}$ yields an order $\mathcal{S}^{\tau P}$ which *still* contains P-worlds, provided $\mathcal{S}$ did.

Table 3 highlights a *typical example* for each class, as introduced earlier in the text (instead of repeating the definition in the table, we merely give a brief reminder).

Observe the inclusion relations among the seven classes: extremely positive attitudes are strongly positive, strongly positive attitudes are positive etc. In terms of our table: the class of attitudes described in row i includes the class of attitudes described in row $i+1$, for $1 \leq i \leq 6$.

Let us now turn to the three new examples of dynamic attitudes mentioned in Table 2.

2.5.1. Moderate Trust. A natural operation on plausibility orders is an operation which we shall call "upwards refinement":[7] given a plausibility order $\mathcal{S}$, we scan $\mathcal{S}$ for pairs of worlds w and v such that $w \in P$ and $v \notin P$. For each such pair that we find, we delete the pair (v, w) from the order, making w strictly more plausible than v. Otherwise, we preserve the original order. We thus split all cells of equiplausible worlds into a P-part and a non-P-part, making the P-part better than the non-P-part. Figure 7 illustrates what is going on in

[7] Upwards refinement was introduced under the name "refinement" by Papini (2001).

a diagram.

Proceeding to a formal definition, the *upwards refinement* $\mathtt{up}$ is the dynamic attitude defined by setting $\mathrm{S}^{\mathtt{up}P} = \mathrm{S}$, and stipulating, for any $w, v \in \mathrm{S}$, that $w \leq_{\mathcal{S}^{\mathtt{up}P}} v$ iff

— ($w \in P$ iff $v \in P$) and $w \leq_{\mathcal{S}} v$, *or*
— $w \in P$, $v \notin P$ and $w \leq_{\mathcal{S}} v$, *or*
— $w \notin P$, $v \in P$ and $w <_{\mathcal{S}} v$.

Rephrasing the above: this simply means that for each $w \in \mathrm{S}$, we split the set $\{v \in \mathrm{S} \mid v \approx_{\mathcal{S}} w\}$ in two, making the P-worlds strictly better than the non-P-worlds, while otherwise preserving the order.

Notice that the fixed point of upwards refinement $\mathtt{up}$ is the propositional attitude *refinedness*, denoted by R and given by

$$\mathcal{S} \vDash RP \text{ iff } \forall w, v \in \mathrm{S}: \text{ if } w \in P \text{ and } v \notin P, \text{ then } w <_{\mathcal{S}} v \text{ or } v <_{\mathcal{S}} w.$$

In terms of upwards refinement $\mathtt{up}$ and minimal trust $\uparrow$, we will define a dynamic attitude which we call *moderate trust*, denote with $\Uparrow$, and which will turn out to be intermediate in strength between minimal trust $\uparrow$ and strong trust $\Uparrow$.[8]

To do so, we first introduce another operation on dynamic attitudes, called the "composite". A pair of dynamic attitudes (σ, τ) is *composable* if the family of upgrades $\{\sigma P \cdot \tau P\}_{P \subseteq W}$ is a dynamic attitude. The *composite* of two dynamic attitudes σ and τ is given by

$$\mathcal{S}^{\sigma \cdot \tau P} := \begin{cases} \mathcal{S}^{\sigma P \cdot \tau P} & (\sigma, \tau) \text{ is composable} \\ \varnothing & \text{otherwise} \end{cases}$$

Now: *moderate trust* $\Uparrow$ is the composite of $\mathtt{up}$ and minimal trust $\uparrow$, that is:

$$\Uparrow := \mathtt{up} \cdot \uparrow$$

Observe that upwards refinement and minimal trust are composable, so moderate trust arises simply by composing the upgrades given by upwards refinement and minimal trust, i.e., $\Uparrow P = \mathtt{up}P \cdot \uparrow P$, for any $P \subseteq W$. Observe also that $\mathtt{up} \cdot \uparrow = \uparrow \cdot \mathtt{up}$; the order does not matter.

We notice here that beliefs induced by performing an upgrade $\Uparrow P$ are *more robust* than beliefs induced by performing an upgrade $\uparrow P$. Consider: after upgrading an order $\mathcal{S}$ with $\Uparrow P$, the agent will continue to believe that

[8] Moderate trust has first been considered under the name of "restrained revision" by Booth and Meyer (2006).

P after performing an additional upgrade with any Q such $\text{best}_{\mathcal{S}\Uparrow P} Q \cap P \neq \varnothing$. This is not the case for minimal trust $\uparrow$! In terms of the subsumption order, we have $\Uparrow < \uparrow$, as can be seen by noting that the fixed point of $\Uparrow$, which is the disjunction of opposite knowledge and refined belief, entails the fixed point of $\uparrow$ (which is the disjunction of opposite knowledge and simple belief). We will discuss moderate trust in greater detail in Chapter 4.

2.5.2. Weak Semi-Trust. While moderate trust strengthens minimal trust, one may also want to consider dynamic attitudes that fail to create belief, but create weaker propositional attitudes instead. An example of such a dynamic attitude is what we call weak semi-trust $\Uparrow^{\sim}$. Consider Figure 8: essentially, applying $\Uparrow^{\sim} P$ to an order $\mathcal{S}$ amounts to making the best P-worlds as good as the worst non-P-worlds, while otherwise preserving the pairs given by $\mathcal{S}$. If there are no P-worlds, or no non-P-worlds, nothing changes, i.e., in that case, $\mathcal{S}^{\Uparrow^{\sim} P} = \mathcal{S}$.

2.5.3. Bare Semi-Trust. An even weaker dynamic attitude is given by *bare semi-trust* $!^{\sim}$, defined as follows:[9]

$$\mathcal{S}^{!^{\sim} P} := \begin{cases} \mathcal{S} & \mathcal{S} \cap P \neq \varnothing, \\ \varnothing & \text{otherwise.} \end{cases}$$

Thus, while an upgrade $\Uparrow^{\sim} P$ on a plausibility order $\mathcal{S}$ (as defined in §2.5.2 right above) may lead to non-trivial changes in $\mathcal{S}$, an upgrade $!^{\sim} P$ amounts to merely *testing* whether there are P-worlds in S. Notice that, indeed, $!^{\sim}$ is a test in the sense of §1.7.4: $!^{\sim}$ is the test for the dual of irrecovable knowledge, $K^{\sim}$, given (recall §1.2.10) by

$$\mathcal{S} \vDash K^{\sim} P \text{ iff } \mathrm{S} \cap P \neq \varnothing.$$

We can interpret $!^{\sim}$ as a dynamic attitude our agent has towards sources which are incapable of providing her with genuine information, but which are "in tune" with her in the minimal sense that they the agent does not receive information from the source that contradicts that she already knows.

2.5.4. What is Special about the Typical Examples?. In the remainder of this section, we give a first answer to the question what is "special" about the typical examples of (the classes of) dynamic attitudes we have identified? One answer uses our notion of a *fixed point* (§1.7):

[9] An analogous notion was introduced by Veltman (1996) to give a semantics for the epistemic modal *might*. We return to this in Chapter 6.

	Semantic Condition	*Characterization*	*Typical Example*
Positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$.	–	Minimal trust $\uparrow$
Strictly Positive	$\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq P$ and if $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$.	Strictures of positive attitudes	Strict minimal trust $\uparrow^{+}$
Negative	If $\mathrm{S} \cap \neg P \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq \neg P$.	Opposites of positive attitudes	Minimal distrust $\uparrow^{\neg}$
Strictly negative	$\mathrm{best}\,\mathcal{S}^{\tau P} \subseteq \neg P$ and if $\mathrm{S} \cap \neg P \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$.	Opposites of strictly positive attitudes	Strict minimal distrust $\uparrow^{-}$
Semi-positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \cap P \neq \varnothing$.	–	Semi-trust $\uparrow^{\sim}$
Semi-negative	If $\neg P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \neq \varnothing$ and $\mathrm{best}\,\mathcal{S}^{\tau P} \cap \neg P \neq \varnothing$.	Opposites of semi-positive attitudes	Semi-distrust $\uparrow^{\neg\sim}$

TABLE 1. Classes of dynamic attitudes related to simple belief.

	Definition of class	*Strictures create*	*Typical example*
1. Irrevocably positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$ and $\mathcal{S}^{\tau P} \subseteq P$	irrevocable knowledge K	infallible trust !
2. Strongly positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$ and for all $w, v \in \mathrm{S}^{\tau P}$: if $w \in P$, $v \notin P$, then $w <_{\mathcal{S}^{\tau P}} v$.	strong belief Sb	strong trust $\Uparrow$
3. Moderately positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$, best $\mathcal{S}^{\tau P} \subseteq P$ and for all $w, v \in \mathrm{S}^{\tau P}$: if $w \in P, v \notin P$ and $w \leq_{\mathcal{S}} v$, then $w <_{\mathcal{S}^{\tau P}} v$.	refined belief Rb	moderate trust $\Uparrow\!\!\!\uparrow$
4. Positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$ and best $\mathcal{S}^{\tau P} \subseteq P$	simple belief B	minimal trust $\uparrow$
5. Semi-positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathcal{S}^{\tau P} \neq \varnothing$ and best $\mathcal{S}^{\tau P} \cap P \neq \varnothing$	dual belief $B^{\sim}$	semi-trust $\uparrow^{\sim}$
6. Weakly semi-positive	If $P \cap \mathrm{S} \neq \varnothing$, then (1) $\mathcal{S}^{\tau P} \neq \varnothing$ and (2) if $\mathcal{S}^{\tau P} \cap \neg \neq \varnothing$, then exist $w, v \in \mathcal{S}^{\tau P}$: $w \in P$, $v \notin P$: $w \leq_{\tau P} v$	dual strong belief $Sb^{\sim}$	weak semi-trust $\Uparrow^{\sim}$
7. Barely positive	If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} \cap P \neq \varnothing$	dual knowledge $K^{\sim}$	bare semi-trust $!^{\sim}$

TABLE 2. Seven Qualitative Degrees of Trust and Semi-Trust.

	Reminder for Definition	*Fixed point of stricture*
1. Infallible trust !	"Delete all non-P-worlds"	The fixed point of $!^{+} = !$ is irrevocable knowledge K
2. Strong trust $\Uparrow$	"Make all P-worlds strictly more plausible than all non-P-worlds"	The fixed point of $\Uparrow^{+}$ is strong belief Sb
3. Moderate Trust $\Uparrow\Uparrow$	"Compose minimal trust with refinement"	The fixed point of $\Uparrow\Uparrow^{+}$ is refined belief Rb
4. Minimal trust $\uparrow$	"Make the best P-worlds the best worlds overall"	The fixed point of $\uparrow^{+}$ is simple belief B
5. Semi-trust $\uparrow^{\sim}$	"Add the best P-worlds to the best worlds overall"	The fixed point of $\uparrow^{\sim +}$ is dual belief ("plausibility") $B^{\sim}$
6. Weak semi-trust $\Uparrow^{\sim}$	"Make the best P-worlds at least as good as the worst non-P-worlds"	The fixed point of $\Uparrow^{\sim +}$ is dual strong belief ("remote plausibility") $Sb^{\sim}$
7. Bare semi-trust	"Test whether there are any P-worlds"	The fixed point of $!^{\sim +} = !^{\sim}$ is dual knowledge ("possibility") $K^{\sim}$

TABLE 3. Seven Typical Examples of Dynamic Attitudes.

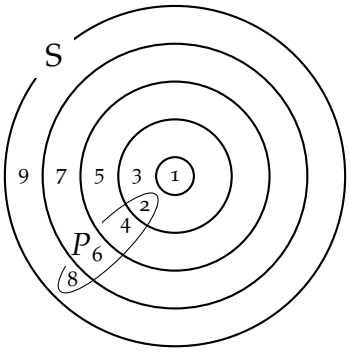


FIGURE 7. Upwards Refinement up. Applying an upgrade up$P$ amounts to breaking all ties between P-worlds and non-P-worlds in favour of the P-worlds, while leaving the order otherwise unchanged.

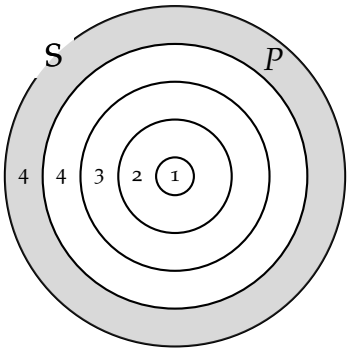


FIGURE 8. Weak semi-trust $\Uparrow^{\sim}$: applying an upgrade $\Uparrow^{\sim} P$ amounts to adding the best P-worlds to the best worlds overall (in the diagram, the P-worlds are given by the gray slice).

PROPOSITION 30.

1. $\overline{!} = K$ *(the fixed point of infallible trust is irrevocable knowledge).*
2. $\overline{\uparrow^{+}} = B$ *(the fixed point of strict minimal trust is belief).*
3. $\overline{\Uparrow^{+}} = Sb$ *(the fixed point of strict strong trust is strong belief).*
4. $\overline{\Uparrow\Uparrow^{+}} = Rb$ *(the fixed point of strict moderate trust is refined belief).*
5. $\overline{\uparrow^{\sim+}} = B^{\sim}$ *(the fixed point of strict semi-trust is the dual of belief).*
6. $\overline{\Uparrow^{\sim+}} = Sb^{\sim}$ *(the fixed point of strict weak semi-trust is the dual of strong belief).*

7. $\overline{!^{\sim}} = K^{\sim}$ *(the fixed point of bare semi-trust is the dual of knowledge).*

So, strict minimal trust, to take an example, is a special example of a strictly positive dynamic attitude because it not only *creates* belief but is also *stopped by* belief (cf. §1.7.2).

A second answer may be given in terms of the notion of *subsumption* (§1.8). Observe that our examples are neatly lined up along the subsumption order:[10]

PROPOSITION 31.

1. $\varnothing < ! < \Uparrow < \Uparrow\!\Uparrow < \uparrow < \uparrow^{\sim} < \Uparrow^{\sim} < !^{\sim} < id.$

2. $\varnothing < ! < \Uparrow^{+} < \Uparrow\!\Uparrow^{+} < \uparrow^{+} < \uparrow^{\sim +} < \Uparrow^{\sim +} < !^{\sim} < id.$

PROOF. We give just one example. To show that $\Uparrow < \Uparrow\!\Uparrow$, we need to verify that for any plausibility order $\mathcal{S}$ and proposition P, $(\mathcal{S}^{\Uparrow P})^{\Uparrow\!\Uparrow P} = \mathcal{S}^{\Uparrow P}$. Assuming that $P \cap \mathrm{S} = \varnothing$, it follows that $\mathcal{S}^{\Uparrow P} = \mathcal{S} = \mathcal{S}^{\Uparrow\!\Uparrow P}$, so our claim holds. Suppose, then, that $P \cap \mathrm{S} = \varnothing$, Then for any $w, v \in \mathrm{S}$ such that $w \in P$, $v \notin P$, we have that $w <_{\mathcal{S}^{\Uparrow P}} v$. In other words: there are no ties between P-worlds and non-P-worlds in $\mathcal{S}^{\Uparrow P}$. Hence $(\mathcal{S}^{\Uparrow P})^{\Uparrow\!\Uparrow P} = \mathcal{S}^{\Uparrow P}$. Again, our claim holds, and the proof is complete. ⊣

Notice that it is not the case that $\Uparrow < \uparrow^{+}$: to see this, one only needs to consider a plausibility order $\mathcal{S}$ and a proposition P such that $P \cap \mathrm{S} = \varnothing$. In such a situation, $\mathcal{S}^{\Uparrow P} = \mathcal{S}$, however, $\mathcal{S}^{\uparrow^{+} P} = \varnothing$, hence $(\mathcal{S}^{\Uparrow P})^{\uparrow^{+} P} \neq \mathcal{S}^{\Uparrow P}$, so $\Uparrow \not< \uparrow^{+}$. It is also not the case that, generally, $\tau^{+} < \tau$ (counterexample: consider $\tau = \uparrow^{+}$; then $\tau^{+} = \tau$. But, quite obviously, it is not the case that $\tau < \tau$).

There is, however, more to say here: our examples are also among the *weakest* in their class, as the following proposition shows (it is essential that in the first and last item we work with τ^{+} rather than τ: otherwise, our argument does not go through).

PROPOSITION 32.

1. *For all extremely positive attitudes* τ: $\tau^{+} \leq !$.

2. *For all strongly positive attitudes* τ: $\tau \leq \Uparrow$.

3. *For all moderately positive attitudes* τ: $\tau \leq \Uparrow\!\Uparrow$.

4. *For all positive attitudes* τ: $\tau \leq \uparrow$.

5. *For all semi-positive attitudes* τ: $\tau \leq \uparrow^{\sim}$.

[10] Observations analogous to the ones in Proposition 31 above can be made for the *opposites* of the dynamic attitudes mentioned in the Proposition.

6. *For all weakly semi-positive attitudes τ: $\tau \leq \Uparrow^{\sim}$.*

7. *For all barely semi-positive attitudes τ: $\tau^{+} \leq !^{\sim}$.*

PROOF. All items are very similar, so we confine ourselves to the first three.

1. Let τ be an extremely positive attitude. Let $\mathcal{S}$ be a plausibility order and P a proposition. Since τ^{+} is strictly extremely positive, $\mathcal{S}^{\tau^{+}P} \vDash KP$. Since the fixed point of ! is K, $(\mathcal{S}^{\tau^{+}P})^{!P} = \mathcal{S}^{\tau^{+}P}$. So $\tau^{+} \leq !$.

2. Let τ be a strongly positive attitude. Let $\mathcal{S}$ be a plausibility order and P a proposition. Since τ is strongly positive, $\mathcal{S}^{\tau P} \vDash (K^{\neg} \vee Sb)P$. Since the fixed point of $\Uparrow$ is $K^{\neg} \vee Sb$, $(\mathcal{S}^{\tau P})^{\Uparrow P} = \mathcal{S}^{\tau P}$. So $\tau \leq \Uparrow$.

3. Let τ be a moderately positive attitude. Let $\mathcal{S}$ be a plausibility order and P a proposition. Since τ is moderately positive, $\mathcal{S}^{\tau P} \vDash K^{\neg}P \vee RbP$. Since the fixed point of $\Uparrow\Uparrow$ is $K^{\neg} \vee Rb$, $(\mathcal{S}^{\tau P})^{\Uparrow\Uparrow P} = \mathcal{S}^{\tau P}$. So $\tau \leq \Uparrow\Uparrow$. ⊣

The above results give us reason to think that our typical examples of dynamic attitudes are indeed "special", as each of them stands in a close connection to a *propositional* attitude of fundamental importance; moreover, they have a special significance as each being among the *weakest* dynamic attitudes in one of the classes we have identified. However, our typical examples are not unique in these two respects, i.e., there are other dynamic attitudes with the same fixed point that are also among the weakest dynamic attitudes in the respective classes. To reuse an earlier example (cf. §2.5.4), consider the test for irrecovable knowledge $?K$, which, by definition (cf. §1.7.4), is given by

$$\mathcal{S}^{?KP} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash KP, \\ \varnothing & \text{otherwise.} \end{cases}$$

The fixed point of $?K$ is obviously irrecovable knowledge K, which is also the fixed point of infallible trust !: $\overline{?\overline{K}} = \bar{!}$. Hence, by Theorem 11, $?K \approx !$ (i.e., $?K$ and ! mutually subsume each other, cf. §1.8.2). Thus, by Proposition 32 above, for any extremely positive attitude τ: $\tau \leq ?K$.

In this sense, $?K$ comes out just as "special" (or "non-special") as !. Why then, do we find it natural to think of ! as *the* dynamic atittude corresponding to irrevocable knowledge K? What makes the connection between infallible trust ! and irrevocable knowledge K *unique*? Analogous questions may be asked for $\Uparrow$, $\uparrow$ and our other typical examples, so the issue clearly deserves further attention. We will discuss it at length in the context of our discussion of *minimal change* in Chapter 3.

2.6. *Mixtures of Dynamic Attitudes*

Our discussion so far has downplayed an important aspect. Realistically, agents rarely assess sources in a completely uniform way: rather, sources are trusted in some contexts, distrusted in others, and perhaps semi-trusted on yet other occasions.

— *Tom Cruise explains what it's like to be a star.*
— *Tom Cruise explains the correct attitude to spiritual life.*

It is perfectly conceivable that some agents might consider Tom Cruise to be an authority on stardom, but ignore his opinions about spiritual life: how such an agent transforms her beliefs upon receiving information from Tom Cruise then depends on the particular topic of conversation.

In such a scenario, none of the classes of dynamic attitudes we have discussed so far offers an appropriate choice. To be able to make such more fine-grained distinctions, what is intuitively needed is a way to *mix* dynamic attitudes. Such mixtures are the topic of the current section.

2.6.1. Mixtures. Mixtures of dynamic attitudes arise by making the choice of a particular attitude towards a source dependent on some feature of the context in which a proposition P is received. In our simple setting, the "context" is represented by the current plausibility order of the agent. As we have seen, static features of epistemic states may be captured by means of propositional attitudes. This leads to the idea of allowing agents to "mix" two dynamic attitudes σ and τ contingent on whether the current epistemic state supports a particular propositional attitude A.[11] Formally, we implement this idea as follows:

— Given two upgrades u and v, an introspective propositional attitude A, and a proposition P, we define the upgrade $u_{AP}v$ by means of

$$\mathcal{S}^{u_{AP}v} := \begin{cases} \mathcal{S}^u & \mathcal{S} \vDash AP \\ \mathcal{S}^v & \mathcal{S} \nvDash AP \end{cases}$$

[11] This is the most natural way to set up a context-dependent notion of trust in our setting, as we allow to "mix" dynamic attitudes using all the information available to the agent that is explicitly modeled in our setting. In extensions of the setting presented here, one might choose to explicitly model further contextual features. For example, one might want to maintain a track record of past information received from a source; or one might want to explicitly model areas of competence of various sources. In such an extended setting, it would be natural to "mix" dynamic attitudes relative to additional parameters of this kind.

— A pair of attitudes (σ, τ) is *A-mixable* if the family of upgrades $\{\sigma P_{AP} \tau P\}_{P \subseteq W}$ is an attitude.[12]
— Given dynamic attitudes σ and τ, and a propositional attitude A, we define the *mixture* $\sigma_A \tau$ *of* σ *and* τ *over* A by means of

$$\mathcal{S}^{(\sigma_A \tau)P} := \begin{cases} \mathcal{S}^{\sigma P_{AP} \tau P} & (\sigma, \tau) \text{ is } A\text{-mixable} \\ \varnothing & \text{otherwise} \end{cases}$$

If σ and τ are A-mixable, then we call $\sigma_A \tau$ a *pure mixture*. Pure mixtures are completely determined by their "components" σ and τ (informally speaking: they contain "no other ingredient" than just σ and τ): if a mixture $\sigma_A \tau$ is pure, then by the above definition, we have that $\mathcal{S}^{\sigma_A \tau P} \in \{\mathcal{S}^{\sigma P}, \mathcal{S}^{\tau P}\}$ for any $\mathcal{S}$ and P.

2.6.2. Example. Let us consider an example of a mixture. Suppose our agent trusts his own eyes unless he is believes he is drunk.[13] Let Q be the set of possible worlds where the agent is drunk. We define the propositional attitude D (for "drunk") by requiring, for any plausibility order $\mathcal{S}$ and proposition Q, that

$$\mathcal{S} \vDash DP \text{ iff } \mathcal{S} \vDash BQ.$$

So the proposition P that D takes an argument is actually ignored: all that matters for DP to satisfied in a plausibility order $\mathcal{S}$ is whether the agent believes he is drunk in $\mathcal{S}$. Suppose that if the agent does not believe that he is drunk, his attitude towards his own eyes is given by minimal trust, while when he is drunk, he ignores what he sees, so his attitude is then given by doxastic neutrality. Then the attitude of the agent towards his own eyes may be given by $\uparrow_{\neg D} id$, the mixture of minimal trust $\uparrow$ and neutrality id over $\neg D$ (the complement of D).

In the same style, one may capture more elaborate cases, for example, an agent who trusts (in the sense of $\uparrow$) his own eyes when he believes he is sober, semi-trusts (in the sense of $\uparrow^{\sim}$) his own eyes when he is not sure whether he is sober or drunk, and ignores what he sees when he believes himself to be drunk.

2.6.3. Pure Mixtures. The next proposition answers the question just which mixtures are pure.

proposition 33. *A pair of attitudes* (σ, τ) *is A-mixable iff for any order* $\mathcal{S}$*, one of the following holds:*

[12] To unpack this definition using the previous line, simply put $u = \sigma P$ and $v = \tau P$.

[13] For the sake of the argument, let us assume that the agent does not "forget" his attitude when drunk; that is: if he is actually drunk, he really does not trust his own eyes.

— $\mathcal{S} \vDash AP$ *and* $\mathcal{S}^{\sigma P} \vDash (AP \vee \overline{\tau} P)$, or
— $\mathcal{S} \vDash \neg AP$ *and* $\mathcal{S}^{\tau P} \vDash (\neg AP \vee \overline{\sigma} P)$.

PROOF. From left to right, suppose that (σ, τ) is A-mixable. Let $\mathcal{S}$ be a plausibility order, and let P be a proposition. We first assume that $\mathcal{S} \vDash AP$. It follows that $\mathcal{S}^{(\sigma_A \tau)P} = \mathcal{S}^{\sigma P}$. If $\mathcal{S}^{\sigma P} \vDash AP$, our claim holds, so suppose that $\mathcal{S}^{\sigma P} \nvDash AP$. Towards a contradiction, suppose that also $\mathcal{S} \nvDash \overline{\tau} P$. Then $(\mathcal{S}^{\sigma P})^{\sigma_A \tau P} = (\mathcal{S}^{\sigma P})^{\tau P} \neq \mathcal{S}^{\sigma P}$. Hence $(\mathcal{S}^{\sigma_A \tau P})^{\sigma_A \tau P} \neq \mathcal{S}^{\sigma_A \tau P}$. Thus the family of upgrades $\{\sigma_A \tau P\}_{P \subseteq W}$ is not a dynamic attitude. So (σ, τ) is not A-mixable. But this contradicts the initial assumption. We conclude that $\mathcal{S} \vDash \overline{\tau} P$, so our claim holds. Second, we assume that $\mathcal{S} \vDash \neg AP$, and argue analogously. This concludes the left to right direction.

From right to left, suppose that the condition given in the statement of the proposition holds. We have to show that $\{\sigma P_{AP} \tau P\}_{P \subseteq W}$ is an attitude. We show that $\sigma P_{AP} \tau P$ is idempotent. Pick an order $\mathcal{S}$ and a proposition P. Suppose that $\mathcal{S} \vDash AP$. Then $\mathcal{S}^{\sigma P_{AP} \tau P} = \mathcal{S}^{\sigma P}$. If $\mathcal{S}^{\sigma P} \vDash AP$, we are done, since σP is idempotent. If $\mathcal{S}^{\sigma P} \nvDash AP$, by the assumption, $\mathcal{S} \vDash \overline{\tau} P$, hence $(\mathcal{S}^{\sigma P})^{\sigma P_{AP} \tau P} = (\mathcal{S}^{\sigma P})^{\tau P} = \mathcal{S}^{\sigma P}$, which shows the claim. Under the assumption that $\mathcal{S} \nvDash AP$, we argue analogously. ⊣

2.6.4. MIXTURES OVER TOPICS. As announced earlier, mixtures allow us to capture context-dependent forms of trust that depend on the "topic" of the information received. Returning to our initial example: some people may consider Tom Cruise to be trustworthy on questions concerning the experience of being a star, but less so on spiritual matters. Similarly, a famous mathematician could be considered very trustworthy when she is making mathematical statements, but less so on administrative matters.

In both cases, whether the source is trusted depends on the topic the information received from the source is about. To a first approximation, we can represent a topic as a set of propositions Γ. Given two dynamic attitudes σ and τ, we would then like to define a dynamic attitude υ such that

$$\mathcal{S}^{\upsilon P} := \begin{cases} \mathcal{S}^{\sigma P} & P \cap \mathrm{S} \in \Gamma, \\ \mathcal{S}^{\tau P} & \text{otherwise.} \end{cases}$$

This means that if P is "on topic" we use σ, and if P is "off topic" we use τ.

To capture this as a mixture, we define the propositional attitude $\check{\Gamma}$ by putting

$$\mathcal{S} \vDash \check{\Gamma} P \text{ iff } P \cap \mathrm{S} \in \Gamma.$$

So $\check{\Gamma}$ simply checks whether P is on topic (given the current order $\mathcal{S}$). We now observe that the mixture of σ and τ over $\check{\Gamma}$ is just υ.

Consider the example of a source who is a mathematician, trusted on "mathematical" propositions, but less so on other matters. More specifically, we may assume that our agent intends to perform an upgrade $\Uparrow P$ whenever receiving a "mathematical" proposition P, but intends to keep her plausibility order unchanged when receiving a "non-mathematical" proposition from that source. Suppose that the set Γ collects all "mathematical" propositions. Then we may capture this dynamic attitude by means of the mixture $\Uparrow_{\check{\Gamma}} id$ (the mixture of strong trust $\Uparrow$ and neutrality id over $\check{\Gamma}$). This can be seen as a "mixed" form of trust.

2.6.5. Further Examples. Mixtures are a versatile tool to define a variety of dynamic attitudes. We discuss a number of further examples.

1. *A variant of minimal trust:* First, consider an agent who, when receiving the information that P from a particular source, comes to believe that P only if she does not yet believe the opposite. If the agent already believes $\neg P$, the incoming information is ignored. In other words, the source only has an effect on the epistemic state of the agent if the agent does not already have an opinion on P. What we have in mind is the attitude τ given by

$$\mathcal{S}^{\tau P} := \begin{cases} \mathcal{S}^{\uparrow P} & \mathcal{S} \not\models B\neg P, \\ \mathcal{S} & \text{otherwise.} \end{cases}$$

   This dynamic attitude can be captured by a mixture, namely: $\tau = \uparrow_{B\neg} id$.

2. *Prioritization:* Let σ and τ be dynamic attitudes. We are interested in capturing an agent that, when receiving information from a particular source, is committed to applying the attitude σ *as long as this is consistently possible*, more precisely, as long as applying σ does not yield an inconsistent epistemic state. Otherwise, she will use the attitude τ. That is, the agent's "overall" attitude can be described by υ, given by

$$\mathcal{S}^{\upsilon P} := \begin{cases} \mathcal{S}^{\sigma P} & \mathcal{S}^{\sigma P} \neq \emptyset, \\ \mathcal{S}^{\tau P} & \text{otherwise.} \end{cases}$$

   To capture υ as a mixture, we define the propositional attitude $\checkmark$ ("executability") by means of

$$\mathcal{S} \models \checkmark_\sigma P \text{ iff } \mathcal{S}^{\sigma P} \neq \emptyset;$$

   we define the *prioritization* $\sigma \ll \tau$ of σ over τ by means of

$$\sigma \ll \tau := \sigma_{\checkmark_\sigma} \tau;$$

   and we notice that, indeed, $\upsilon = \sigma \ll \tau$.

3. Recall (§1.7.4) that for any propositional attitude A, the *test for* A, denoted by $?A$, is the dynamic attitude given by

$$\mathcal{S}^{?AP} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash AP, \\ \varnothing & \text{otherwise.} \end{cases}$$

   We now observe that tests can be captured by mixtures, indeed, for any A, we have $?A = id_A\varnothing$.

4. *Restrictions:* The information an agent can receive from a source may be *limited* in the sense that the agent is only able to consistently process information if she already has a particular propositional attitude to the information received; otherwise, she will end up in the inconsistent epistemic state. We have in mind a dynamic attitude σ which, given another dynamic attitude τ and a propositional attitude A, is defined by

$$\mathcal{S}^{\sigma P} := \begin{cases} \mathcal{S}^{\tau P} & \mathcal{S} \vDash AP \\ \varnothing & \text{otherwise} \end{cases}$$

   This may be captured by a mixture, which we call the *restriction of* τ *to* A, denoted τ_A, and defined by $\tau_A := \tau_A\varnothing$. Obviously, $\sigma = \tau_A$.

5. *Strictures:* As a special case of a restriction, we can recover the notion of a stricture. So far, we have defined strictures for positive attitudes only; however, one easily defines, for any dynamic attitude τ, the stricture τ^+ of τ by means of

$$\mathcal{S}^{\tau^+ P} = \begin{cases} \mathcal{S}^{\tau P} & \mathcal{S} \cap P \neq \varnothing, \\ \varnothing & \text{otherwise,} \end{cases}$$

   The stricture of τ is just the restriction of τ to $K^\sim$: $\tau^+ = \tau_{K^\sim}$. So strictures are restrictions, and thus mixtures.

2.7. Mutual Assessments of Reliability

In this section, we move towards a more encompassing modeling style. Besides evidence given by sense data, or derived in an inferential manner ("indirect" evidence), an agent typically obtains information *from other agents*. While our framework provides the resources to formalize this aspect (as we will see in this chapter), we have not made it explicit so far. In the setting developed in the previous sections, sources of information are featureless parameters, individuated only by the trust (or lack of trust) an (implicit) agent places in

them. They are external to the formal model. As we aspire to model testimony provided to agents by other agents, it is natural to model the epistemic state not only of the *recipient*, but also of the *sender*. This makes the source a part of the model, and will allow us to capture a variety of characteristic properties of testimony.

The multi-agent version of our setting presented in this section (and illustrated with a number of examples in the next one) meshes well with the strengths of dynamic epistemic logic, a research tradition that historically originated with the aim of providing a detailed account of the informational dynamics of interaction among communicating agents.[14] The work of this section contributes directly to this line of research.

2.7.1. Multi-Agent Plausibility Orders. Fix a finite, non-empty set $\mathcal{A}$ (the *set of agents*). A multi-agent plausibility order (over $\mathcal{A}$) is a family of agent-indexed preorders (i.e., reflexive, transitive relations)

$$\{\mathcal{S}_a := (\mathrm{S}, \leq^a_{\mathcal{S}})\}_{a\in\mathcal{A}},$$

containing one preorder $\mathcal{S}_a$ on S for each agent $a \in \mathcal{A}$.[15]

A multi-agent plausibility order $\{\mathcal{S}_a := (\mathrm{S}, \leq^a_{\mathcal{S}})\}_{a\in\mathcal{A}}$ differs from a single-agent plausibility order $\mathcal{S}$ (cf. §1.1.1) in two respects: first, and most obviously, a multi-agent plausibility order $\{\mathcal{S}_a := (\mathrm{S}, \leq^a_{\mathcal{S}})\}_{a\in\mathcal{A}}$ gives us a *collection* of preorders $\mathcal{S}_a$, one for each agent $a \in \mathcal{A}$. Second, notice that the members of this family of preorders are *not* (as one might have expected) single-agent plausibility orders. Rather, they are only required to be reflexive and transitive. To see more clearly why this is conceptually reasonable, let us explore the structure given by a multi-agent plausibility order in more detail.

2.7.2. Information Cells. Given a multi-agent plausibility order $(\{\mathcal{S}_a\}_{a\in\mathcal{A}}, [\![\cdot]\!])$, let, for each agent a, the relation $\sim^a_{\mathcal{S}}$ be given by:

$$w \sim^a_{\mathcal{S}} v \text{ iff } w \leq^a_{\mathcal{S}} v \text{ or } v \leq^a_{\mathcal{S}} w.$$

The relation $\sim^a_{\mathcal{S}}$ captures epistemic indistinguishability: The fact that $w \sim^a_{\mathcal{S}}$ indicates that at the world w, the actual world could just as well be v, for

[14] Cf. the early references Plaza (1989), Gerbrandy (1999), Baltag et al. (1999), whose main concern was to formally capture the dynamics of information flow in multi-agent scenarios. A focus on interaction has remained a trademark of the field, as evidence, e.g., by the title of van Benthem's recent monograph (van Benthem 2011).

[15] Notice that the preorders share the same domain S.

all agent a (irrevocably) knows. Put differently, agent a does not have hard information at w that would allow him to exclude that the actual world is v.[16]

Now for each world $w \in \mathrm{S}$, let $a_{\mathcal{S}}(w) := \{v \in \mathrm{S} \mid w \sim^a_{\mathcal{S}} v\}$. We call $a_{\mathcal{S}}(w)$ the *information cell* of a at w in $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$. It represents the agent's hard information at w (in $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$). Put differently, $a_{\mathcal{S}}(w)$ indicates how the world w *appears* to agent a, capturing all the "epistemic alternatives" of w, all the worlds that could be the actual world according to the agent's hard information at w. Since $\sim^a_{\mathcal{S}}$ is, clearly, an equivalence, it follows that whenever $w \sim^a_{\mathcal{S}} v$, then $a_{\mathcal{S}}(w) = a_{\mathcal{S}}(v)$.

2.7.3. Local States. Each information cell $a_{\mathcal{S}}(w)$ induces an ordering given by

$$\mathcal{S}_{a(w)} := \mathcal{S}_a \cap (a_{\mathcal{S}}(w), a_{\mathcal{S}}(w) \times a_{\mathcal{S}}(w)).$$

We call $\mathcal{S}_{a(w)}$ the *local state* of agent a at w in $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$.

The domain of the local state $\mathcal{S}_{a(w)}$ consists of the information cell $a_{\mathcal{S}}(w)$ of agent a at w; and the relational pairs in the local state $\mathcal{S}_{a(w)}$ are given by those $(w, v) \in \mathcal{S}_a$ such that $w, w \in a_{\mathcal{S}}(w)$.

Besides the "appearance" of the world w to agent a (given by $a_{\mathcal{S}}(w)$), a local state also captures the "plausibility hierarchy" imposed on all the worlds that are consistent with the agent's hard information at w.

Observe that, by the above definition of $a_{\mathcal{S}}(w)$ in terms of $\sim^a_{\mathcal{S}}$, which is in turn defined in terms of $\leq^a_{\mathcal{S}}$, the order $\mathcal{S}_{a(w)}$ is reflexive, transitive and connected, so $\mathcal{S}_{a(w)}$ is, in fact, a (single-agent) plausibility order in the sense of §1.1.

Notice that this plausibility order, the local state $\mathcal{S}_{a(w)}$, presents the "appearance" of w to agent a viewed in isolation from other agents. If other agents are taken into account, then $\mathcal{S}_{a(w)}$ may turn out, in a sense, "too small." The other agents may have hard information differing from the hard information of agent a; in particular, agent a may not know exactly what hard information the other agents have! Hence the need for our notion of a "bigger" structure in which the agents' local states live. In this way, multi-agent plausibility orders also represent the agents' uncertainty about each other. This also explains why the preorders $\mathcal{S}_a$ that live in a multi-agent plausibility order $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ are not required to be connected. Requiring connectedness would amount to stipulating that all agents have *the same* hard information! But this is clearly unreasonable.

[16] Since $\sim^a_{\mathcal{S}}$ is clearly an equivalence relation, this also holds, as is desirable, *vice versa*: by symmetry of $\sim^a_{\mathcal{S}}$, at the world v, the actual world could just as well be w, for all agent a knows. Of course, the following are also true: by reflexivity of $\sim^a_{\mathcal{S}}$, at the world w, the actual world could just as well be w, for all agent a knows. And the same for world v.

As we have seen, connectedness does reappear on the level of local states of agents. Since the local state of a single agent is just a single-agent plausibility orders, the notion of a local state provides a link that ties the multi-agent structures we work with in this section together with our previous work on single-agent plausibility orders: all earlier definitions and results apply directly to local states, and thus indirectly to multi-agent plausibility orders.

2.7.4. Example. For illustration of the preceding concepts, we provide an example of a multi-agent plausibility order in Figure 9. The information cells for each agent in $\{a, b\}$ in this example are given by

— $a_{\mathcal{S}}(x) = \{x\}$, $b_{\mathcal{S}}(x) = \{x, w, y\}$,
— $a_{\mathcal{S}}(y) = \{w, y\}$, $b_{\mathcal{S}}(y) = \{x, w, y\}$,
— $a_{\mathcal{S}}(w) = \{w, y\}$, $b_{\mathcal{S}}(w) = \{x, y, w\}$.

So while at any of the three worlds w, x and y, agent b is not able to exclude any of the other two from consideration, agent a has hard information that the actual world is x at world x, while at both worlds w and y, agent a is uncertain whether the actual world is w or y.

As detailed above, each information cell gives rise to a local state that encompasses both the hard information (given by the cell) and the soft information (given by the plausibility order restricted to the cell) of the agent at the worlds in the cell. For example, the local state $\mathcal{S}_{a(y)}$ of agent a induced by the cell $a_{\mathcal{S}}(x)$ is given by the single-agent plausibility order

$$\mathcal{S}_{a(y)} = (a_{\mathcal{S}}(y), \{(y, w), (w, y), (w, w), (y, y)\}),$$

while $\mathcal{S}_{b(y)}$ (the local state of agent b induced by the cell $a_{\mathcal{S}}(y)$) is given by the single-agent plausibility order

$$\mathcal{S}_{b(y)} = (a_{\mathcal{S}}(y), \{(w, x), (w, y), (x, y), (y, x), (w, w), (x, x), (y, y)\}).$$

So at world y, agent a has hard information that $\{w, y\}$ is satisfied, and he considers both worlds w and y to be equiplausible. Agent b, on the other hand, has hard information that $\{w, x, y\}$ is satisfied, among which he considers w to be the most plausible ones, with x and y equiplausible to each other, but strictly less plausible than w.

2.7.5. Trust Graphs. We are now interested in capturing how each agent a assesses the reliability of each other agent b. As we will see below, this assessment will determine how a upgrades her plausibility order upon receiving

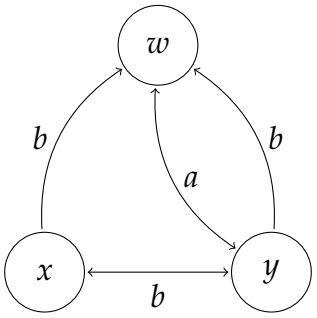


FIGURE 9. Example of a multi-agent plausibility order for the set of agents $\mathcal{A} = \{a, b\}$. Reflexive and transitive loops are omitted.

information from b. We collect the mutual attitudes of the agents in a structure that we call a *trust graph*.

A *trust graph* is a function T that assigns to each pair of agents a and b such that $a \neq b$ a dynamic attitude $T(a, b)$.

Here, the fact that $T(a, b) = \tau$ is interpreted as indicating that agent a has the dynamic attitude τ towards agent b.

Writing $(a, \tau, b) \in T$ iff $T(a, b) = \tau$ brings out clearly why we choose to call T a trust graph: we think of the agents as the nodes, and the sources as the labels of a labeled graph, which can be drawn in the familiar way. Figure 10 gives an example of such a drawing. In the diagram, the $\uparrow$-edge originating from a going to c indicates, for example, that agent a has the attitude minimal trust towards agent b (according to the trust graph T defined by the diagram). Similarly, according to the diagram, agent a has the attitude $\Uparrow_{\check{\Gamma}} id$ (the mixture of strong trust and neutrality over $\check{\Gamma}$, with Γ representing a particular topic, cf. §2.6.4) towards agent b.

Notice that, by virtue of T being a function, the following properties are generally satisfied:

1. For any two agents $a, b \in \mathcal{A}$ such that $a \neq b$: there exists a dynamic attitude τ such that $(a, \tau, b) \in T$ (*existence*).

2. For any two agents $a, b \in \mathcal{A}$ such that $a \neq b$: if $(a, \tau, b) \in T$ and $(a, \tau', b) \in T$, then $\tau = \tau'$. (*uniqueness*).

This brings out our implicit assumption that each agent assesses the reliability of each other agent *in a particular, unique way*. This assumption is not entirely without substance. "I don't know whether to trust Peter" is a sensible statement, and in making such a statement, one could be taken to express that

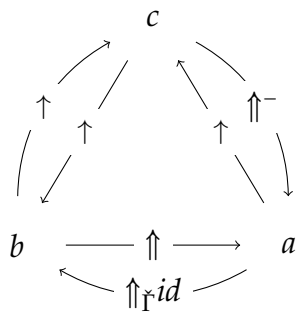


FIGURE 10. An example of a trust graph, with the nodes given by agents, and the directed edges given by dynamic attitudes that represent mutual assessments of reliability. For example, the dynamic attitude of agent a towards agent c is given by minimal trust $\uparrow$ according to the example.

one does not assess Peter's reliability in a particular, unique way. But it is a bit unclear what the statement would correspond to on the level of our formal models. Does it mean that the agent who makes the statement cannot decide which dynamic attitude to apply when she receives information from Peter? Surely, she will change her information state in *some* particular (unique) way when receiving information from Peter, so another way of reading the statement would be to presume that the speaker's dynamic attitude towards Peter is characterized by a certain ambiguity, so that certainly information from Peter does not induce belief, but perhaps only the dual of belief (cf. §2.4). If that is what statements like the above mean, then our formalism handles them without problem.

Rather than taking a definitive stance on the issue, we simply set it aside: in the following, we work on the assumption that our agents *have* made up their mind, and thus, the above *existence* and *uniqueness* requirements are satisfied.

2.7.6. TRUST-PLAUSIBILITY ORDERS. Consider a multi-agent plausibility order $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ as defined in §2.7.1 above. We would like to formally encode some additional information, answering the question how each agent $b \in \mathcal{A}$ assesses the reliability of each other agent $c \in \mathcal{A}$ (where $c \neq b$). For this, our notion of a trust graph comes in handy. However, we need a slightly richer concept. While we shall assume that agents are introspective about how reliable they consider other agents to be, they might very well be uncertain about the reliability assessments of these other agents. Hence, we need to allow that

trust graphs vary across the possible worlds in S. This leads to the following definition.

Let $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ a multi-agent plausibility order. A *trust labeling* over $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ is a function T assigning to each possible world in W a trust graph T_w, satisfying for each agent $b, c \in \mathcal{A}$:

$$\text{If} \in T_w(b,c) = \tau \text{ and } w \sim_b v, \text{ then } T_v(b,c) = \tau.$$

So a trust labeling T gives us a trust graph T_w for each possible world $w \in$ S, with the additional requirement statet above expressing that agents are introspective about their own dynamic attitudes towards other agents: the dynamic attitudes of agent a towards other agents do not vary within a's information cell at w.

A *(multi-agent) trust-plausibility order* is a pair

$$(\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$$

where $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ is a multi-agent plausibility order, and T is a trust labeling over $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$.

2.7.7. Trust-Plausibility Transformer. A *trust-plausibility transformer* $[c]$ is a function

$$\mathcal{C} \overset{[c]}{\longmapsto} \mathcal{C}[c]$$

assigning a trust-plausibility order

$$\mathcal{C}[c] := (\{\mathcal{S}_a[c] := (\mathrm{S}[c], \leq^a_{\mathcal{S}_a[c]})\}_{a\in\mathcal{A}}, T[c])$$

to each given trust-plausibility order $\mathcal{C} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, in such a way that $\mathrm{S}[c] \subseteq$ S and $T[c] = T$.

A trust-plausibility transformer is thus a way of transforming the representation of a joint information state given by a trust-plausibility model $\mathcal{C}$ into a new joint information state, given by the trust-plausibility model $\mathcal{C}[c]$.

Given a world $w \in$ S, the trust-plausibility transformer $[c]$ is *executable* in $\mathcal{C}$ at w iff $w \in \mathrm{S}[c]$.

Notice that to uniquely determine a trust-plausibility transformer $[c]$, it is enough to specify, for each trust-plausibility order $(\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, what $[c]$ does to the underlying multi-agent plausibility order $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$, since the effect of $[c]$ on T is *fixed* by the general definition, i.e., T is just copied into the new structure. So formally specifying the order $\mathcal{S}_a[c]$, for each agent a and for each given trust-plausibility order $(\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, defines a communication act $[c]$.

Keeping the trust labeling fixed on application of a trust-plausibility transformer encodes the assumption that agents do not change their mutual assessments of realiability as new information comes in. This is not a realistic assumption, of course. It is perfectly conceivable that an agent receives information that makes her, for example, lose trust in another agent. Modeling the *dynamics of trust* in our framework is a topic we leave for future research.[17]

2.7.8. Communication Acts. The above notion of a trust-plausibility transformer is rather generic in that it does not identify the trigger of a particular transformation. The idea of a communication act is to consider specific trust-plausibility transformers that are triggered by an agent communicating a piece of information to the other agents. Communication acts will thus specify *who* is making the communication act, and precisely *what* proposition is communicated in the communication act.

To arrive at our desired notion, we will exploit the resources given to us by the trust labeling which is part of a trust-plausibility order: whenever an agent a makes a communication act, we assume that the other agents *apply* their dynamic attitudes towards a (as given by the trust labeling) to upgrade their information states. In other words: they implement their strategy for belief change.

More concretely, given an agent $b \in \mathcal{A}$, and a proposition P, we want to define a trust-plausibility transformer $[b{:}\, P]$, which we will call a communication act. So for each given trust-plausibility order $\mathcal{C} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, agent $b \in \mathcal{A}$, and proposition $P \subseteq W$, we would like to specify a multi-agent plausibility order $\{\mathcal{S}_a[b{:}\, P]\}$. As observed right above, doing this defines a unique trust-plausibility transformer.

What intuitively needs to be done is roughly this: we need to upgrade, for each agent $a \in \mathcal{A}$ and each possible world $w \in \mathrm{S}$, the local state $\mathcal{S}_{a(w)}$ according to the dynamic attitude of agent a towards b; and then we need to collect all the upgraded orders in a single structure.

Given a trust-plausibility order $\mathcal{C} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, [\![\cdot]\!], L)$ and a world $w \in \mathrm{S}$, we make the notation

$$\tau^{w}_{a\to b} := T_w(a, b)$$

Start now with the new domain $\mathrm{S}[b{:}\, P]$. It will be given by the worlds

[17] An extension of our framework in this direction could profit from existing work on the dynamics of trust in the multi-agent systems literature, cf., e.g., Falcone and Castelfranchi (2004), Boella and van der Torre (2005). For further remarks on the dynamics of trust, see the conclusion of this dissertation.

surviving the appropriate upgrade for each agent $a \neq b$, that is:

$$\mathrm{S}[b\colon P] := \{w \in \mathrm{S} \mid \forall a \neq b \in \mathcal{A} : w \in (\mathcal{S}_{b(w)})^{\tau^w_{a\to b}P}\}.$$

We write $\mathcal{S}[b\colon P]$ for the natural product order on $\mathrm{S}[b\colon P]$, that is,

$$\mathcal{S}[b\colon P] := (\mathrm{S}[b\colon P], \mathrm{S}[b\colon P] \times \mathrm{S}[b\colon P]).$$

To obtain the orderings on the new domain $\mathrm{S}[b\colon P]$ for each agent b such that $b \neq a$, we take—as announced above—the union of the "individually upgraded" local states (as described above) and intersect the latter union with the product order $\mathcal{S}[b\colon P]$. So for any $a \neq b$, put:

$$\mathcal{S}_a[b\colon P] := \big(\bigcup_{w\in\mathrm{S}} (\mathcal{S}_{b(w)})^{\tau^w_{a\to b}P}\big) \cap \mathcal{S}[b\colon P]$$

For agent b, we simply put

$$\mathcal{S}_b[b\colon P] := \mathcal{S}_b \cap \mathcal{S}[b\colon P],$$

i.e., we restrict the old order $\mathcal{S}_b$ to the new domain $\mathcal{S}[b\colon P]$.

Let us sum up and write down a formal definition. For every agent $b \in \mathcal{A}$ and proposition $P \subseteq W$, the *communication act* $[b\colon P]$ is given by the trust-plausibility transformer that assigns to each given trust-plausibility model $\mathcal{C}$ the trust-plausibility model

$$\mathcal{C}[b\colon P] := (\{\mathcal{S}_a[b\colon P]\}_{a\in\mathcal{A}}, T),$$

given by:

$$\mathrm{S}[b\colon P] := \{w \in \mathrm{S} \mid \forall a \neq b \in \mathcal{A} : w \in (\mathcal{S}_{b(w)})^{\tau^w_{a\to b}P}\},$$

where we assume that $T_w(a,b) = \tau^w_{a\to b}$, and

$$\forall a \neq b : \mathcal{S}_a[b\colon P] := \big(\bigcup_{w\in\mathrm{S}} (\mathcal{S}_{b(w)})^{\tau^w_{a\to b}P}\big) \cap \mathcal{S}[b\colon P],$$

where, recall from above, $\mathcal{S}[b\colon P]$ is the product order on $\mathrm{S}[b\colon P]$, and

$$\mathcal{S}_b[b\colon P] := \mathcal{S}_b \cap \mathcal{S}[b\colon P].$$

2.7.9. Example: Indirect Learning from a Source. We illustrate the notion of a communication act by means of an example that brings out a distinctive feature of our setting. Consider the trust graph depicted in Figure 12, and the multi-agent plausibility order given by Figure 11. In this example, we assume that the mutual assessments of reliability among the agents are common knowledge. That is, we assume that the trust graph depicted in Figure

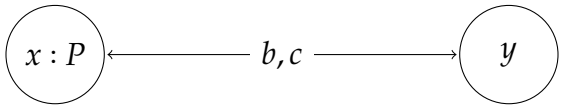


FIGURE 11. Example of a multi-agent plausibility order for the set of agents $\{a, b, c\}$. The labeling of the drawing indicates that $x \in P$, while it is not the case that $y \in P$. Both agents b and c consider both worlds equiplausible. Reflexive loops are omitted. Agent a can epistemically distinguish between the two worlds, so no arrows are drawn for this agent.

12 is associated with *both* worlds in the order given by 11. So taken together the two diagrams define a trust-plausibility order $\mathcal{C}$.

Now consider what happens to this trust-plausibility order when the communication act $[a{:}\, P]$ is made. Upon first inspection, one might expect that agent b will respond by making the world x more plausible than the world y in her plausibility order, since she strongly trusts the speaker a, as given by the fact that $T_w(b, a) = \Uparrow$. However, notice that the attitude of agent c to agent b is given by infallible trust !: this means that the world y will *not* be an element of $\mathcal{S}_{c(y)}$. By our definition of a communication act, y will thus *not* be contained in $\mathcal{C}[a{:}\, p]$. So in fact, $\mathcal{C}[a{:}\, p]$ will be the trust-plausibility model built over the singleton world x!

This is an example of *indirect learning*: since agent b obtains the *hard information* that p is the case, all other agents, and in particular agent c, will *also* obtain this information. The reason for this is that agent c *knows* that agent b *knows* that information obtained from a comes with a warranty of truthfulness (as captured by infallible trust !). Thus even though c does not have evidence of her own that would guarantee a's trustworthiness, she can rely on the fact that b has such evidence, as indicated by the trust graph.

2.8. Epistemic Norms of Communication

In our single-agent setting, we have studied how incoming information changes the information state of a single agent in a way that depends on the agent's assessment of the reliability of the source of information. Since sources were not taken to be agents, but rather remained anonymous, living outside our formal models, there was not much that could be said about them, except that they happened to be sources our agent has a particular dynamic attitude towards. The multi-agent setting introduced in the previous section is different. In this framework, properties of communication acts made by an agent

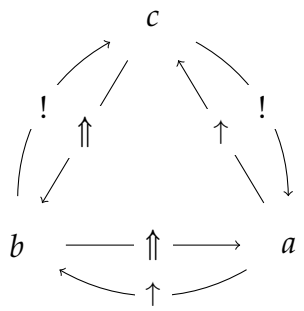


FIGURE 12. Another example of a trust graph.

come into view. In particular, the setting invites normative assessments of communication acts. An agent a may be subjected to various epistemic norms pertaining to the communication acts $[a{:}\, P]$ performed by a. In this section, we illustrate how such norms can be formulated and investigated in our setting by means of a number of examples.

Epistemic norms are often understood as norms which circumscribe the conditions under which it is epistemically permissible to hold certain beliefs.[18] Here, we use the term in a different sense: from the present perspective, epistemic norms circumscribe the conditions under which a certain communication act (representing an assertion made by an agent) is "permissible", given mutual assessments of reliability within a group of agents, and assuming certain (pre-formal) standards of how a trustworthy agent would behave. Agents might, for example, be subjected to the requirement of saying what they believe, not inducing false beliefs in others, disclosing all relevant information, and so forth. The question is how such requirements can be captured formally. Our purpose in this section is not to argue for any specific epistemic norm; we simply aim to illustrate how our setting can be used to analyze some examples of possible norms, norms that one may or may not want to impose on agents and their communication acts.

Our approach is to formulate epistemic norms as *properties of communication acts*; in the following, we will consider two such norms: *sincerity* and *honesty*. While sincerity relates communication acts to the speaker's belief, honesty takes into account the attitude of the hearer toward the speaker. Having introduced this pair of notions, we first establish that there is a tension between the two. Then, we relate them to what we take to be a fundamental interest of a speaker in an information exchange, namely that she will gener-

[18] Cf. Pollock (1987).

ally aim at *persuading* hearers w.r.t. certain propositions, getting them to adopt the same attitude towards the proposition she already has herself.[19]

2.8.1. Preliminaries. Start with some notation and terminology.

In the following, given a communication act $[a{:}\,P]$, we often refer to agent a as "the speaker", and to any other agent b (with $b \neq a$) as "a hearer."

Given a trust-plausibility order $\mathcal{C} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, an introspective propositional attitude A, a proposition P, an agent $a \in \mathcal{A}$ and a world $w \in \mathrm{S}$, let us write

$$\mathcal{C}, w \vDash A_a P \text{ iff } \mathcal{S}_{a(w)} \vDash AP.$$

Because of its familiarity from epistemic logic, this notation is self-explanatory. We say that the agent a has the propositional attitude A towards P in $\mathcal{C}$ at w iff $\mathcal{C}, w \vDash A_a P$.

For the remainder of this section, fix a trust-plausibility order $\mathcal{C} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T)$, propositions P and Q, an agent $a \in \mathcal{A}$ ("the speaker") and a world $w \in \mathrm{S}$.

2.8.2. Sincerity. We say that the communication act $[a{:}\,P]$ is *sincere in $\mathcal{C}$ at w* iff

$$\mathcal{C}, w \vDash B_a P.$$

So sincerity requires that the agent believes what she is asserting, i.e., $[a{:}\,\varphi]$ is sincere in a trust-plausibility order at a world w if the agent simply believes P in $\mathcal{C}$ at w (which, by the notation introduced above, is the same as saying that she believes P in her local state $\mathcal{S}_{a(w)}$ at w in $\mathcal{C}$).

This is about the most simple formal notion of sincerity in our setting. There are other reasonable ones. For instance, a notion of "strong sincerity" may be obtained by replacing simple belief with strong belief in the preceding definition. But for our purposes, the definition above, in terms of simple belief, will suffice.

2.8.3. Honesty. We say that the communication act $[a{:}\,P]$ is *honest towards agent b in $\mathcal{C}$ at w* iff

$$\text{if } T_w(b, a) = \tau, \text{ then } \mathcal{C}, w \vDash \overline{\tau}_a P.$$

We say that $[a{:}\,P]$ is *honest in $\mathcal{C}$ at w* iff $[a{:}\,P]$ is honest towards every $b \neq a \in \mathcal{A}$ in $\mathcal{C}$ at w.

[19] The work of this section builds on Baltag and Smets (in print). The notion of sincerity is also important in the context of the logical analysis of lying (van Ditmarsch, van Eijck, Sietsma, and Wang 2012). Persuasiveness is also studied in the multi-agent systems literature (Dunin-Kęplicz and Verbrugge 2010) and in argumentation theory (Walton and Krabbe 1995).

The intuition behind the notion of honesty is that a speaker should not induce propositional attitudes in a hearer she does not have herself. In view of this intuition, it is natural to require, in order for the communication act $[a{:}\,P]$ to be honest, that the propositional attitude of the speaker a towards P should be matched by the dynamic attitude of a hearer b towards a. Here, we exploit the fact that the notion of a fixed point introduced in §1.7 allows us to connect dynamic and propositional attitudes.

2.8.4. Example. As an example, consider the communication act $[a{:}\,P]$ and suppose $T_w(b,a) = \uparrow^+$, i.e., the attitude of agent b towards agent a at w is strict minimal trust $\uparrow^+$. As we know, the fixed point of strict minimal trust $\uparrow^+$ is simple belief B (Proposition 30). For $[a{:}\,P]$ to be honest at w it is then required that $\mathcal{C}, w \vDash B_a P$.

Since dynamic attitudes create their fixed point, we have, assuming that $w \in \mathrm{S}[a{:}\,P]$, that $\mathcal{C}[a{:}\,P], w \vDash BP$. So the propositional attitude towards P of the speaker before the communication act is matched by the propositional attitude of the hearer after the communication act. In this way, our notion of honesty reflects the intuition cited above: a speaker should not induce propositional attitudes in a hearer she does not have herself.

2.8.5. The Responsibility of Honest Speakers. The following observations are immediate consequences of earlier results:

proposition 34. 1. *If* $T_w(b,a) = !$, *then* $[a{:}\,P]$ *is honest towards* b *at* w *in* $\mathcal{C}$ *iff* $\mathcal{C}, w \vDash K_a P$.

2. *If* $T_w(b,a) = \Uparrow^+$, *then* $[a{:}\,P]$ *is honest towards* b *at* w *in* $\mathcal{C}$ *iff* $\mathcal{C}, w \vDash Sb_a P$.

3. *If* $T_w(b,a) = \uparrow^+$, *then* $[a{:}\,P]$ *is honest towards* b *at* w *in* $\mathcal{C}$ *iff* $\mathcal{C}, w \vDash B_a P$.

proof. We show the first item. Suppose that $T_w(b,a) = !$. Then $[a{:}\,P]$ is honest iff $\mathcal{C}, w \vDash \bar{!}_a P$. But since $\bar{!} = K$, the latter is the case iff $\mathcal{C}, w \vDash K_a \varphi$. ⊣

In words: asserting what you know (resp.: what you strongly believe, resp.: what you believe) is a necessary and sufficient condition for being honest towards an agent who infallibly trusts you (resp.: strongly positively trusts you, resp.: minimally positively trusts you).

This highlights the fact that the requirements imposed on an honest speaker are, in our formalization, not given in absolute terms, but relative to the level of trust bestowed upon the speaker by a hearer.

This seems to reflect an important feature of our everyday conception of trust: the more people trust you, the higher your responsibility in carefully

"weighing your words". Your audience might just take the truth of what you say for granted. This is different from our notion of sincerity: sincerity just requires the speaker to "speak his mind", i.e., say what she believes.

2.8.6. The Tension between Honesty and Sincerity. Honesty and sincerity capture two plausible epistemic norms of communication, namely "be honest!" (the norm of honesty) and "be sincere!" (the norm of sincerity). There is, however, a tension between the two norms: in certain circumstances, it is *impossible* to fulfill both of them simultaneously. Consider the following memorable dialogue (taken from the movie *Pirates of the Caribbean*):

> Mullroy: *What's your purpose in Port Royal, Mr. Smith?*
>
> Murtogg: *Yeah, and no lies.*
>
> Mr. Smith (aka Jack Sparrow): *Well, then, I confess, it is my intention to commandeer one of these ships, pick up a crew in Tortuga, raid, pillage, plunder and otherwise pilfer my weasely black guts out.*
>
> Murtogg: *I said no lies.*
>
> Mullroy: *I think he's telling the truth.*
>
> Murtogg: *Don't be stupid: if he were telling the truth, he wouldn't have told it to us.*
>
> Jack Sparrow: *Unless, of course, he knew you wouldn't believe the truth even if he told it to you.*

Suppose our agent a is actually Jack Sparrow, to some people known as "Mr. Smith", who intends, at the world w, to pick up a crew in Tortuga, pillage, plunder and otherwise pilfer his weasely black guts out, and suppose that P is the set of possible worlds where he has these intentions (so in particular $w \in P$). Suppose further that our agent b is actually Murtogg, who does not trust agent a. In fact, let us assume that the dynamic attitude of b towards a at the current world of evaluation, w, is given by a strictly negative attitude (cf. §2.3). Let us also suppose that Jack Sparrow is fully introspective about his own intentions, and assume that *a irrecovably knows* that P at w.

In these circumstances, the communication act $[a{:}\, p]$ is sincere at w: Sparrow just speaks his mind, revealing his true intentions. However, $[a{:}\, p]$ is not honest. Since agent b has a negative attitude towards agent a, say τ, the fixed point of τ entails the opposite of belief $B^{\neg}$, hence for $[a{:}\, P]$ to be honest, it is required that $\mathcal{C}, w \vDash B\neg P$. But since $\mathcal{C}, w \vDash KP$ by our assumption, it is not the case that $\mathcal{C}, w \vDash B\neg P$, so $[a{:}\, P]$ is not honest.

On our formalization, then, $[a\colon p]$ constitutes a "sincere lie" in $\mathcal{C}$ at w, fulfilling the norm of sincerity, while violating the norm of honesty. So the sincerity of a communication act does not imply its honesty. The communication act $[a\colon \neg P]$, on the other hand, is honest, but not sincere in $\mathcal{C}$ at w, so honesty does not imply sincerity either.

Let us, for the sake of the argument, assume that Jack Sparrow is actually interested in conveying the truth about his intentions to Murtogg. Our analysis shows that this is easier said than done. This highlights the fact that an atmosphere of distrust is bound to put strains on the integrity of the agents. Assuming that they want to obey the norms of sincerity and honesty, how can they convince others that a certain proposition is satisfied? As the above example shows, the straightforward approach of "speaking your mind" does not always work; and simply saying the opposite of what you believe—when dealing with a distrusting hearer—may not work either. We consider this further in §2.8.8 below.

2.8.7. Honest Minimally Trusted Agents are Sincere. In natural language, the terms "honesty" and "sincerity" seem to be used almost interchangeably. Prima facie, our formalization is at odds with this observation. However, our setting does reflect ordinary use, in a sense, in view of the above-mentioned fact that a minimally trusted agent is honest iff she believes what she is saying iff she is sincere.

More formally, if the attitude of a hearer b towards the speaker a at w is given by strict minimal trust $\uparrow^+$, then $[a\colon P]$ is honest towards b at w iff $[a\colon P]$ is sincere at w, as a consequence of Proposition 34 above.

In fact, the weaker assumption that the attitude of b towards a is a strictly positive dynamic attitude already guarantees that a communication act $[a\colon P$ that is honest at w in $\mathcal{C}$ is also sincere at w in $\mathcal{C}$.

To see this, suppose that $T_w(b,a) = \tau$ is strictly positive, and assume that $[a\colon \varphi]$ is honest towards b at w in $\mathcal{C}$. By definition of honesty, $\mathcal{C}, w \vDash \overline{\tau}_a P$. But our earlier results imply that $\overline{\tau} \leq B$ for any strictly positive τ (*Proof:* Suppose τ is strictly positive. By Proposition 32, $\tau \leq \uparrow^+$. By Proposition 11, $\overline{\tau} \leq \overline{\uparrow^+}$, and since $\overline{\uparrow^+} = B$, it follows that $\overline{\tau} \leq B$.) Thus, since $\mathcal{C}, w \vDash \overline{\tau}_a P$, it follows that $\mathcal{C}, w \vDash B_a P$, so $[a\colon P]$ is sincere at w in $\mathcal{C}$.

It is not, however, in general the case, that a sincere communication act is honest towards a hearer who has a strictly positive atttitude towards the speaker. Honesty will fail if the hearer trusts a "too much." As an example: we know from Proposition 34, that if a hearer b has the attitude $\Uparrow^+$ towards the speaker a, then the speaker needs to *strongly believe* that P for the communication act $[a\colon P]$ to be honest; sincerity, on the other hand, merely guarantees

that the speaker *simply believes* that P.

2.8.8. Persuasiveness. The communication act $[a{:}\, P]$ is *persuasive w.r.t. Q towards b in $\mathcal{C}$ at w* iff

$$\text{if } \mathcal{C}, w \vDash B_a Q, \text{ then } \mathcal{C}[a{:}\, P], w \vDash B_b Q.$$

A communication act $[a{:}\, P]$ is thus persuasive towards b w.r.t. some "issue" Q at w iff the communication act gets the hearer b to adopt a belief after the communication act that the speaker held before.

As with sincerity, one could also define other notions of persuasiveness, replacing, for example, B (for simple belief) with Sb (for strong belief) in the above definition. But again, for our purposes, the simple notion will suffice.

2.8.9. How to Be Persuasive, Sincere and Honest. We may now wonder what it takes for a speaker to persuade a hearer that Q (i.e., get the hearer to believe that Q) using some honest and sincere communication act.

We work with an example. Suppose that agent a is actually George W. Bush who wants to convert agent b, the American people, that there are weapons of mass destruction in Iraq. As it happens, we suppose, agent *a simply believes* himself that there are weapons of mass destruction, but does not *strongly* believe it (because, one might add, his evidence for the existence of said weapons is rather sketchy). Further, we assume, agent *b strictly strongly trusts* agent a. Let us also assume that agent b does *not already* believe that there are weapons of mass destruction in Iraq (otherwise, agent a does not have much persuading to do).

Let Q be the set of worlds where there are weapons of mass destruction in Iraq. Formally, we assume that $\mathcal{C}$ and w are such that

1. $T_w(b, a) = \Uparrow^+$,
2. $\mathcal{C}, w \vDash B_a Q$,
3. $\mathcal{C}, w \vDash \neg Sb_a Q$.
4. $\mathcal{C}, w \vDash \neg B_b Q$,

For illustration, consider the multi-agent plausibility order depicted in Figure 13, which is consistent with the above list of assumptions. But notice that the following argument does not depend on the particulars of this multi-agent plausibility order: we will base the argument just on the four assumptions above.

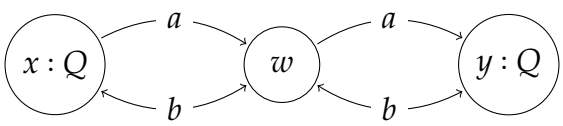


FIGURE 13. An example context for the Bush scenario. Q is the set of worlds in which there are weapons of mass destruction in Iraq, so the labeling indicates that there are such weapons in world x and y, but not in world w.

Our question is: is there a communication act $[a{:}\,P]$ which is honest, sincere and persuasive w.r.t. Q towards b in $\mathcal{C}$ at w? Spelled out in more detail, $[a{:}\,P]$ meets our requirements iff

— $\mathcal{C}, w \vDash B_a P$ (required by sincerity),
— $w \vDash_{\mathcal{C}} Sb_a Q$ (required by honesty), and
— $w \vDash_{\mathcal{C}[a:P]} B_b Q$ (required by persuasiveness).

Let us first consider three options that will *not* work:

1. Merely asserting that Q will not do: $[a{:}\,Q]$, while sincere and persuasive, is *not* honest, since, by our assumption, $\mathcal{C}, w \nvDash Sb_a Q$.

2. Another option to consider for a is to assert that he irrevocably knows that Q. This would amount to choosing P by means of

$$P := \{v \in \mathrm{S} \mid \mathcal{C}, v \vDash K_a Q\}.$$

   But assuming this choice of P, unfortunately, $[a{:}\,P]$ is *neither* sincere *nor* honest. Agent a does not simply believe that he knows that P (as required by sincerity), let alone that he would strongly believes that he knows that P (as required by honesty).

3. A third idea would be for a to assert that he (simply) believes that Q, that is, a could choose P by means of

$$P := \{v \in \mathrm{S} \mid \mathcal{C}, v \vDash B_a Q\}$$

   Under this assumption, $[a{:}\,P]$ is sincere and honest, however, $[a{:}\,P]$ is *not* guaranteed to be persuasive. As a counter-example, consider Figure 13: here, agent b knows that agent a believes that Q, and the same goes, of course, for agent a himself. As a consequence, the communication act $[a{:}\,P]$ applied to $\mathcal{C}$ yields just $\mathcal{C}$: the American people is unimpressed by learning that George W. Bush believes that there are weapons of mass destruction in Iraq.

Interestingly, however, a solution for agent a's problem is available: he can assert that he *defeasibly knows that* Q, that is, choose P by means of

$$P := \{v \in \mathrm{S} \,||\, \mathcal{C}, v \vDash \Box_a Q\}.$$

Let us verify that, under this choice of P, $[a{:}\,P]$ is sincere, honest and persuasive towards b in $\mathcal{C}$ at w w.r.t. Q.

— *Sincerity:* Observe that for any trust-plausibility order $\mathcal{C}'$ and world $v \in \mathrm{S}'$: $\mathcal{C}', v \vDash B_a Q$ iff $\mathcal{C}', v \vDash B_a \Box_a Q$. In other words: agents take their beliefs to be defeasible knowledge. So the communication act $[a{:}\,P]$ is sincere iff the communication act $[a{:}\,Q]$ is sincere. But since $\mathcal{C}, w \vDash B_a Q$, $[a{:}\,Q]$ *is* in fact sincere, so the same goes for $[a{:}\,P]$.

— *Honesty:* for any trust-plausibility order $\mathcal{C}'$ and world $v \in \mathrm{S}'$: $\mathcal{C}', v \vDash B_a Q$ iff $\mathcal{C}', w \vDash Sb_a \Box_a Q$. Since, by our assumption, $\mathcal{C}, w \vDash B_a Q$, it follows that $\mathcal{C}, w \vDash Sb_a \Box_a Q$. For $[a{:}\,Q]$ to be honest, on the other hand, it is required that $\mathcal{C}, w \vDash Sb_a P$ (since $T_w(b, a) = \Uparrow^+$), but this is just what we have verified, so $\mathcal{C}, w \vDash Sb_a P$. So $[a{:}\,P]$ is honest.

— *Persuasiveness:* To see that $[a{:}\,P]$ is persuasive, notice that in $\mathcal{C}[a{:}\,P]$, agent b will consider all worlds in $P \cap \mathrm{S}[a{:}\,P]$ strictly more plausible than all other worlds. This implies that $\mathcal{C}[a{:}\,P], w \vDash B_b \Box_a Q$. But notice that for any trust-plausibility order $\mathcal{C}'$ and world $v \in \mathrm{S}'$: if $\mathcal{C}', v \vDash \Box_a Q$, then $\mathcal{C}', v \vDash Q$. Hence $\mathcal{C}[a{:}\,Q], w \vDash B_b Q$. So $[a{:}\,P]$ is persuasive.

Interestingly, $[a{:}\,P]$ (based on our last choice of P, as just discussed at length) is very close to what agent a (or at least, his close associate, agent c, aka Dick Cheney) actually told agent b, as a matter of historical fact: "We know that there are weapons of mass destruction in Iraq." One outcome of our analysis is that he could not have possibly had irrevocable knowledge in mind when he spoke of knowledge in this context; another is that if we translate "know" as defeasible knowledge $\Box$, then what he said was actually impeccable in view of the norms we have formulated above. How, then, could anyone get the impression that Bush was in violation of moral standards (as some people have claimed)? Perhaps the answer is that a responsible agent will hold himself to a standard of *epistemic transparency*, disclosing the evidence—or lack of evidence—on which his assertions are based. And in Bush's case, the evidence was sketchy. Representing this type of consideration formally is a topic for future research.

Chapter 3. Minimal Change

The concept of *minimal change* is of the first importance in many theories of belief change.[1] In particular, most of the crucial AGM postulates have been motivated by appeal to minimal change.[2] The thesis usually associated with the concept is that in order to revise with a proposition P, one should transform one's epistemic state in such a way as to ensure that one afterwards believes P, but, crucially, in doing so should keep the "difference" to the original epistemic state as small as possible. The plausibility of the thesis derives from the fact that any "non-minimal" change seems to do more than is needed. In this vein, Gilbert Harman suggested the following principle:

> *Principle of Conservatism:* One is justified in continuing to fully accept something in the absence of a special reason not to.[3]

As it stands, the principle of conservatism is too weak to enforce the notion that a rational agent should only minimally modify her epistemic state when accomodating new information. One way to more fully justify this *principle of minimal change* is by combining Harman's principle with considerations of informational economy: one is justified to keep prior doxastic commitments one

[1] Cf., e.g., Harman (1986), Gärdenfors (1988), Arlo-Costa and Levi (2006). Minimal change also plays an important role in the semantic theory of natural language. Lewis's work on counterfactuals relies on the idea of evaluating the consequent of a counterfactual in the closest worlds (to the "actual world") satisfying its antecedent. These worlds are those that *minimally differ* from (are minimally changed compared to) the actual world (while satisfying the antecedent). In work on counterfactuals in the dynamic semantics tradition, operations are studied that allow us to algorithmically determine minimally changed worlds satisfying the antecedent from given ones, cf. Veltman (2005). A related perspective is supplied by causal treatments of counterfactuals inspired by the "structural equations" approach due to Pearl (2000), in particular, cf. Schulz (2011). While this chapter deals with minimal change from the perspective of belief revision theory, our results might be relevant for formal semantics as well.

[2] Cf. Alchourrón et al. (1985).

[3] Harman (1986).

is not forced to give up (*principle of conservatism*); but dropping commitments one is justified to keep is a disproportionate response, overly costly at the least (*principle of informational economy*).[4] Hence it is rational to maintain all commitments one is not forced to give up—the principle of minimal change thus follows from Harman's principle plus the principle of informational economy.

Even if one does not subscribe to the principle of minimal change from a philosophical perspective (perhaps on grounds of rejecting informational economy as a universal norm of rationality), the concept of minimal change provides a useful way to evaluate and compare different belief change policies. However, one needs to ask: minimal *for what purpose*? Boutilier called his favourite belief revision method "natural revision," because it seemed to him to appropriately formalize a notion of minimal change, or "conservatism" of belief change.[5] Darwiche and Pearl, on the other hand, have criticized Boutilier's method, arguing, essentially, that it produces beliefs that are not "robust" enough under further revision.[6] This type of dissent points to the fact that what counts as a minimal change should be evaluated in view of what revision *aims at*: it is the minimal change required to meet some target condition. If the aim of revision is, simply, to acquire *belief* (in the formal sense given by the propositional attitude B) in the proposition received from a source, Boutilier's proposal is a very plausible candidate for an optimal choice. If, on the other hand, the epistemic state resulting from revision is required to satisfy additional constraints, other policies might be required that minimally change given orders so as to *meet those constraints*.[7]

The proposal of this chapter is thus to adopt a *flexible measure of optimality* that is sensitive to the target of revision, i.e., the propositional attitude towards the information received that the revision is meant to achieve. In our setting, belief change policies will be given by dynamic attitudes; and they will count as optimal if they reach their specific target in a minimal way. What is that target? Naturally, for each dynamic attitude τ, we shall take its target as given by the fixed point $\overline{\tau}$ of τ, i.e., the propositional attitude realized by τ (cf. §1.7).

[4] Board (2004) puts the point like this: "our beliefs are not in general gratuitous, and so when we change them in response to new evidence, the change should be no greater than is necessary to incorporate that new evidence."

[5] Cf. Boutilier (1993). In this dissertation, "natural revision" corresponds to minimal trust $\uparrow$.

[6] Cf. Darwiche and Pearl (1996).

[7] A related perspective is provided by Baltag, Gierasimczuk, and Smets (2011), whose work shows that adopting natural revision/minimal upgrade as a revision policy significantly restricts the capacity of an agent to *learn* given structures in the long-term. So, again, if additional constraints (such as, in this case, the *long-term* goal of learning a structure) are put in place, minimal upgrade may be an inappropriate choice.

The *minimal change problem* then becomes the problem of characterizing which (if any) dynamic attitudes induce their fixed points in an optimal way.

3.1. Similarity

Roughly speaking, we will call a dynamic attitude τ "optimal" for its fixed point $\overline{\tau}$ if no other dynamic attitude realizes *the same* fixed point while changing given input orders *less*; and a dynamic attitude τ will be called "canonical" if it is *uniquely optimal* (i.e., τ is the *only* attitude that meets the criterion for optimality).

Our first task is thus to settle on an appropriate notion of "closeness", or "similarity" between plausibility orders.

3.1.1. Similarity. In this chapter, and the next one, we set the multi-agent setting introduced in §2.7 aside and focus on the single-agent case.[8]

Let $\mathcal{S}$ and $\mathcal{S}'$ be two (single-agent) plausibility orders. Thinking of $\mathcal{S}'$ as the plausibility order resulting from $\mathcal{S}$ due to the application of some upgrade, we can compare the two by the extent to which they agree on the relative plausibility of given pairs of worlds.

Formally, suppose that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$ (cf. §2.4.6 for the notation), i.e., there exists some upgrade u such that $\mathcal{S}^u = \mathcal{S}'$, or, equivalently: $S' \subseteq S$.

We say that $\mathcal{S}$ and $\mathcal{S}'$ *agree on* $(w, v) \in S' \times S'$ iff

$$(w, v) \in \mathcal{S} \text{ iff } (w, v) \in \mathcal{S}'.$$

The *agreement set of* $\mathcal{S}$ *and* $\mathcal{S}'$ is the set of pairs (w, v) such that $\mathcal{S}$ and $\mathcal{S}'$ agree on (w, v). Introducing notation:

$$agree_{\mathcal{S}}\mathcal{S}' := \{(w, v) \in S' \times S' \mid \mathcal{S} \text{ and } \mathcal{S}' \text{ agree on } (w, v)\}.$$

For any plausibility order $\mathcal{S}$, the *(strict) similarity order* $(O_{\mathcal{S}}, <_{\mathcal{S}})$ is then defined in the following way. First, $\mathcal{S}' \in O_{\mathcal{S}}$ iff $S' \subseteq S$ (i.e., iff $\mathcal{S} \twoheadrightarrow \mathcal{S}'$). Second, for any $\mathcal{S}', \mathcal{S}'' \in O_{\mathcal{S}}$:

$$\mathcal{S}' <_{\mathcal{S}} \mathcal{S}'' \text{ iff } S'' \subset S' \text{ or } (S'' = S' \text{ and } agree_{\mathcal{S}}\mathcal{S}'' \subset agree_{\mathcal{S}}\mathcal{S}').$$

If $\mathcal{S}' <_{\mathcal{S}} \mathcal{S}''$, then we say that $\mathcal{S}'$ is *more similar to* $\mathcal{S}$ *than* $\mathcal{S}''$ (or that $\mathcal{S}'$ *is closer to* $\mathcal{S}$ *than* $\mathcal{S}''$).

Notice that this definition has a "lexicographic" component: it favours keeping *elements* of the original domain S over preserving *pairs* in the relation

[8] We will return to the multi-agent setting only towards the end of Chapter 5.

$\leq_{\mathcal{S}}$. Recall that S represents the *hard information* of the agent, while $\leq_{\mathcal{S}}$ represents her *soft information*. The intuition is that an (irrevocable) increase in hard information should make more of a difference than a (defeasible) change in soft information. As an example, suppose that $\mathcal{S}$, $\mathcal{S}'$ and $\mathcal{S}''$ are given as follows:

$$\mathcal{S} = (\{w, v\}, \{(w, v), (w, w), (v, v)\}),$$

$$\mathcal{S}' = (\{w\}, \{(w, w)\}),$$

$$\mathcal{S}'' = (\{w, v\}, \{(v, w), (w, w), (v, v)\}).$$

Then $\mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}'$, since $\mathrm{S}' \subset \mathrm{S}''$.

Going further, for any plausibility order $\mathcal{S}$, the *weak similarity order* $(O_{\mathcal{S}}, \leq_{\mathcal{S}})$ is defined by: $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$ iff either $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$ or $\mathcal{S}' = \mathcal{S}''$. More explicitly, this amounts to the following:

LEMMA 35. *$\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$ iff* either $\mathrm{S}' \supset \mathrm{S}''$ or *($\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}''$).*

PROOF. From left to right, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$ implies by definition that $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$ or $\mathcal{S}' = \mathcal{S}''$. In the first case, $\mathrm{S}' \supset \mathrm{S}''$ or $\mathrm{S} = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' \supset agree_{\mathcal{S}}\mathcal{S}''$. In the second case, $\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' = agree_{\mathcal{S}}\mathcal{S}''$. In either case, $\mathrm{S}' \supset \mathrm{S}''$ or ($\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}''$), which completes one half.

For the other half, suppose that $\mathrm{S}' \supset \mathrm{S}''$ or ($\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}''$). First, if $\mathrm{S}' \supset \mathrm{S}''$, then $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$, so $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$. Second, if $\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' \supset agree_{\mathcal{S}}\mathcal{S}''$, then also $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$, so $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$.

Third, suppose that $\mathrm{S}' = \mathrm{S}''$ and $agree_{\mathcal{S}}\mathcal{S}' = agree_{\mathcal{S}}\mathcal{S}''$. We claim that $\mathcal{S}' = \mathcal{S}''$, so $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$, which finishes the proof. To show the claim, consider any pair $(w, v) \in \mathrm{S}' \times \mathrm{S}'$ ($= \mathrm{S}'' \times \mathrm{S}''$). Suppose that $(w, v) \in agree_{\mathcal{S}}\mathcal{S}'$. Then $(w, v) \in \mathcal{S}'$ iff (by definition) $(w, v) \in \mathcal{S}$ iff (by the assumption) $(w, v) \in \mathcal{S}''$. On the other hand, suppose that $(w, v) \notin agree_{\mathcal{S}}\mathcal{S}'$. If $(w, v) \in \mathcal{S}'$, then (by definition) $(w, v) \notin \mathcal{S}$, so (by the assumption) $(w, v) \in \mathcal{S}''$; and analogously: if $(w, v) \notin \mathcal{S}'$, then $(w, v) \notin \mathcal{S}$, so $(w, v) \notin \mathcal{S}''$. The claim holds. The proof is complete. ⊣

3.1.2. PROPERTIES OF THE SIMILARITY ORDER. The weak similarity ordering $(O_{\mathcal{S}}, \leq_{\mathcal{S}})$ has a number of desirable properties that are easy to check:

PROPOSITION 36. *$(O_{\mathcal{S}}, \leq_{\mathcal{S}})$ is a partial order bounded by $\mathcal{S}$ and $\varnothing$. In other words, for any $\mathcal{S}', \mathcal{S}'', \mathcal{S}''' \in O_{\mathcal{S}}$, the following properties are satisfied:*

1. Reflexivity: $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}'$.
2. Transitivity: *If $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$ and $\mathcal{S}'' \leq_{\mathcal{S}} \mathcal{S}'''$, then $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}'''$.*
3. Antisymmetry: *If $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}''$ and $\mathcal{S}'' \leq_{\mathcal{S}} \mathcal{S}'$, then $\mathcal{S}' = \mathcal{S}''$.*

4. Boundedness: *If* $\mathcal{S}' \neq \mathcal{S}$, *then* $\mathcal{S} \prec_{\mathcal{S}} \mathcal{S}'$, *and if* $\mathcal{S}' \neq \varnothing$, *then* $\mathcal{S}' \prec_{\mathcal{S}} \varnothing$.

We can now put the definition of similarity to work by defining suitable notions of optimality and canonicity in terms of it, in the manner sketched above.

3.2. *Optimality*

We shall call a dynamic attitude τ optimal if it realizes its fixed point $\overline{\tau}$ in a way that is minimal compared to any other attitude σ with the same fixed point $\overline{\tau}$. We cash this out by requiring that no such σ is capable of making a "smaller step" along the similarity order. Formally, this yields the following notion of optimality.

3.2.1. Optimality. Let τ be a dynamic attitude. We say that τ is *optimal* iff there is no order $\mathcal{S}$, proposition P and dynamic attitude σ such that

$$\overline{\sigma} = \overline{\tau} \text{ and } \mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\tau P}.$$

Given a propositional attitude A, we say that τ is *optimal for* A if τ is optimal and $\overline{\tau} = A$.

3.2.2. Some Optimal Dynamic Attitudes. Let us first mention some optimal dynamic attitudes.

proposition 37.

1. *Infallible trust* ! *is optimal.*
2. *Strong trust* $\Uparrow$ *is optimal.*
3. *Minimal trust* $\uparrow$ *is optimal.*
4. *Neutrality id is optimal.*
5. *Isolation* $\varnothing$ *is optimal.*

proof.

1. Towards a contradiction, suppose that ! is not optimal. This implies that there exists a plausibility order $\mathcal{S}$, a proposition P, and an attitude σ such that $\overline{\sigma} = K = \overline{!}$ and $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{!P}$. The latter implies that either (1) $\mathrm{S}^{\sigma P} \supset \mathrm{S}^{!P}$ or (2) $\mathrm{S}^{\sigma P} = \mathrm{S}^{!P}$ and $agree_{\mathcal{S}}\mathcal{S}^{\sigma P} \supset agree_{\mathcal{S}}\mathcal{S}^{!P}$. Assuming (1), we observe that $P \cap \mathrm{S} = \mathrm{S}^{!P}$, and since $\mathrm{S}^{\sigma P} \supset \mathrm{S}^{!P}$, it follows that $\mathrm{S}^{\sigma P} \cap \neg P \neq \varnothing$, hence $\mathcal{S}^{\sigma P} \not\models KP$, contradiction. Assuming (2), we derive a contradiction to the fact that $agree_{\mathcal{S}}\mathcal{S}^{!P} = \mathrm{S}^{!P} \times \mathrm{S}^{!P}$. Hence ! is optimal for K.

2. Towards a contradiction, suppose that $\Uparrow$ is not optimal. Then there exists a plausibility order $\mathcal{S}$, a proposition P, and an attitude σ such that $\overline{\sigma} = Sb \vee K^{\neg}$ and $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\Uparrow P}$. If $P \cap \mathrm{S} = \varnothing$, then $\mathcal{S}^{\sigma P} = \mathcal{S}^{\Uparrow P} = \mathcal{S}$, contradiction (using the fourth item of Proposition 36). So $P \cap \mathrm{S} \neq \varnothing$. Also, $\mathrm{S}^{\sigma P} = \mathrm{S}^{\Uparrow P} = \mathrm{S}$, for otherwise, $\mathcal{S}^{\Uparrow P} \prec_{\mathcal{S}} \mathcal{S}^{\sigma P}$, contradiction. So *either* (1) there exist $w, v \in \mathrm{S}$ such that $w \leq_{\mathcal{S}} v$, $w \leq_{\mathcal{S}^{\sigma P}} v$, but not $w \leq_{\mathcal{S}^{\Uparrow P}} v$, *or* (2) there exist $w, v \in \mathrm{S}$ such that not $w \leq_{\mathcal{S}} v$, and not $w \leq_{\mathcal{S}^{\sigma P}} v$, but $w \leq_{\mathcal{S}^{\Uparrow P}} v$. We consider (1) and (2) in turn. Starting with (1), we may distinguish four sub-cases: (a) $w, v \in P$; (b) $w, v \notin P$; (c) $w \in P, v \notin P$; (d) $w \notin P, v \in P$. As for (a) to (c), in each of these sub-cases we immediately find a contradiction with the definition of $\Uparrow$. As for (d), using our assumption and the fact that $\mathcal{S}^{\sigma P} \not\models K^{\neg} P$, we obtain a contradiction with the fact that $\mathcal{S}^{\sigma P} \models SbP$ We continue with (2). From the fact that not $w \leq_{\mathcal{S}} v$ and $w \leq_{\mathcal{S}^{\Uparrow P}} v$, we infer that $w \in P$, $v \notin P$. Thus, from the fact that it is not the case that $w \leq_{\mathcal{S}^{\sigma P}} v$, we infer that $\mathcal{S}^{\sigma P} \not\models SbP$. But, again, we also have $\mathcal{S}^{\sigma P} \not\models K\neg P$, so we have a contradiction to the initial assumption, and this concludes case (2). So $\Uparrow$ is optimal for $Sb \vee K^{\neg}$.

3. Towards a contradiction, suppose that $\uparrow$ is not optimal. Then there exists a plausibility order $\mathcal{S}$, a proposition P, and an attitude σ such that $\overline{\sigma} = B \vee K^{\neg}$ and $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\uparrow P}$. As in the proof of the previous item, we conclude that $P \cap \mathrm{S} \neq \varnothing$ and $\mathrm{S}^{\sigma P} = \mathrm{S}$. So *either* (1) there exist $w, v \in \mathrm{S}$ such that $w \leq_{\mathcal{S}} v$, $w \leq_{\mathcal{S}^{\sigma P}} v$, but not $w \leq_{\mathcal{S}^{\uparrow P}} v$, *or* (2) there exist $w, v \in W$ such that not $w \leq_{\mathcal{S}} v$, and not $w \leq_{\mathcal{S}^{\sigma P}} v$, but $w \leq_{\mathcal{S}^{\uparrow P}} v$. Start with (1). We distinguish four sub-cases: (a) $w, v \in P$; (b) $w, v \notin P$; (c) $w \in P, v \notin P$; (d) $w \notin P, v \in P$. As for (a) to (c), each of these sub-cases immediately yields a contradiction to the definition of $\uparrow$.

   It remains to consider case (d), i.e., we suppose $w \notin P, v \in P$, $w \leq_{\mathcal{S}} v$, $w \leq_{\mathcal{S}^{\sigma P}} v$, while it is not the case that $w \leq_{\mathcal{S}^{\uparrow P}} v$. We conclude by definition of $\uparrow$ that $v \in \mathrm{best}_{\mathcal{S}} P$. We now make three observations:

   — Since $\mathcal{S}^{\sigma P} \models BP$ and $w \leq_{\mathcal{S}^{\sigma P}} v$, it follows that $v \notin \mathrm{best}\, \mathcal{S}^{\sigma P}$. So there exists $x \in \mathrm{best}\, \mathcal{S}^{\sigma P}$ such that $x <_{\mathcal{S}^{\sigma P}} v$.
   — Since $v \in \mathrm{best}_{\mathcal{S}} P$, also $v \in \mathrm{best}\, \mathcal{S}^{\uparrow P}$, hence $v \leq_{\mathcal{S}^{\uparrow P}} x$.
   — Since $v, x \in P$ and $v \in \mathrm{best}_{\mathcal{S}} P$, it follows that $v \leq_{\mathcal{S}} x$.

   Overall, we now have the following situation: $v \leq_{\mathcal{S}} x$, $v \leq_{\mathcal{S}^{\uparrow P}} x$, while it is not the case that $v \leq_{\mathcal{S}^{\sigma P}} x$. This implies that $(v, x) \in \mathrm{agree}_{\mathcal{S}} \mathcal{S}^{\uparrow P}$, $(v, x) \notin \mathrm{agree}_{\mathcal{S}} \mathcal{S}^{\sigma P}$. So it is not the case that $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\uparrow P}$, and this is a contradiction. Since both (1) and (2) yield a contradiction, $\uparrow$ is optimal for B.

4. Immediate by the fact that $\mathcal{S} \models \top$ for any $\mathcal{S}$ and $\mathcal{S} \prec_{\mathcal{S}} \mathcal{S}'$ for any $\mathcal{S}'$ such that

$\mathcal{S} \neq \mathcal{S}'$ (cf. the fourth item of Proposition 36).

5. Immediate by the fact that $\mathcal{S} \models \bot$ implies $\mathcal{S} = \varnothing$. ⊣

Each of these dynamic attitudes thus provides *a* solution to the minimal change problem (relative to its respective fixed point).

Our next objective is to show that, for each introspective propositional attitude A, there is *some* dynamic attitude that is optimal for A. We start by introducing two notions that will be useful throughout the chapter.

3.2.3. Closest Orders. Let A be an introspective propositional attitude. Define

$$\mathrm{opt}_{\mathcal{S}}\, AP := \{\mathcal{S}' \in O_{\mathcal{S}} \mid \mathcal{S}' \models AP, \neg\exists \mathcal{S}'' \in O_{\mathcal{S}} : \mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}', \mathcal{S}'' \models AP\}.$$

Given an input order $\mathcal{S}$, the set $\mathrm{opt}_{\mathcal{S}}\, AP$ captures the set of orders reachable from $\mathcal{S}$ that satisfy AP while differing minimally from $\mathcal{S}$. Notice that $\mathrm{opt}_{\mathcal{S}}\, AP$ is, in general, non-empty:

lemma 38. *For any introspective propositional attitude A, plausibility order $\mathcal{S}$, and proposition P: the set* $\mathrm{opt}_{\mathcal{S}}\, AP$ *is non-empty.*

proof. Choose an introspective propositional attitude A, a plausibility order $\mathcal{S}$, and a proposition P. We know from Theorem 9 that there is a dynamic attitude τ such that $\overline{\tau} = A$. So there exists an order $\mathcal{S}' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \models AP$ and $\mathcal{S} \preceq_{\mathcal{S}} \mathcal{S}'$, namely $\mathcal{S}' = \mathcal{S}^{\tau P}$ (notice that $\mathrm{S}^{\tau P} \subseteq \mathrm{S}$, as required for membership in $O_{\mathcal{S}}$). It follows from the overall finiteness of our setting that $\prec_{\mathcal{S}}$ is well-founded. Thus, given that we have $\mathcal{S}'$, reachable from $\mathcal{S}$ and satisfying AP, we will be able to find an order $\mathcal{S}''$, reachable from $\mathcal{S}$, satisfying AP, and being minimally different from $\mathcal{S}$. Hence $\mathrm{opt}_{\mathcal{S}}\, AP$ is non-empty. ⊣

Notice also that if $\mathcal{S} \models AP$, then $\mathrm{opt}_{\mathcal{S}}\, AP = \{\mathcal{S}\}$, since, by Proposition 36 (item 4), $\mathcal{S} \prec_{\mathcal{S}} \mathcal{S}'$ for any $\mathcal{S}' \neq \mathcal{S}$.

3.2.4. Substantial Propositions. Let $\mathcal{S}$ be a plausibility order. P is *insubstantial in* $\mathcal{S}$ (or *$\mathcal{S}$-insubstantial*) if $\mathrm{S} \cap P \in \{\varnothing, \mathrm{S}\}$. P is *substantial in* $\mathcal{S}$ (or $\mathcal{S}$-substantial) otherwise, i.e., P is $\mathcal{S}$-substantial iff $\varnothing \subset \mathrm{S} \cap P \subset \mathrm{S}$ iff neither $P \cap \mathrm{S} = \varnothing$ nor $P \cap \mathrm{S} = \mathrm{S}$.

As a consequence of Proposition 5, for any dynamic attitude τ, plausibility order $\mathcal{S}$ and proposition P: if P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$. The next lemma will be useful in later sections.

lemma 39. *Suppose $\overline{\sigma} = \overline{\tau}$. If P is $\mathcal{S}$-insubstantial, then $\mathcal{S}^{\sigma P} = \mathcal{S}^{\tau P}$.*

PROOF. Let $\mathcal{S}$ be a plausibility order, and suppose that P is insubstantial in $\mathcal{S}$, i.e., $Q = P \cap \mathrm{S} \in \{\mathrm{S}, \varnothing\}$. Notice that $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau Q}$ by dynamic conservativity (cf. §1.5.6). By Proposition 5, $\mathcal{S}^{\tau Q} \in \{\mathcal{S}, \varnothing\}$. Suppose first that $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau Q} = \mathcal{S}$. Then $\mathcal{S} \vDash \overline{\tau} Q$, and since $\overline{\sigma} = \overline{\tau}$, we conclude that $\mathcal{S}^{\sigma Q} = \mathcal{S}$. Using dynamic conservativity again, $\mathcal{S}^{\sigma Q} = \mathcal{S}^{\sigma P}$. So $\mathcal{S}^{\tau P} = \mathcal{S}^{\sigma P}$. The claim holds. The second case is similar: suppose that $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau Q} = \varnothing$. Then $\mathcal{S} \nvDash \overline{\tau} Q$, so $\mathcal{S}^{\sigma Q} \neq \mathcal{S}$ (since $\overline{\sigma} = \overline{\tau}$), hence $\mathcal{S}^{\sigma Q} = \mathcal{S}^{\sigma P} = \varnothing$ (by Proposition 5), so $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau Q}$, and again, the claim holds. So if P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\sigma P} = \mathcal{S}^{\tau P}$. ⊣

So any attitude τ that realizes a particular propositional attitude A (in the sense that $\overline{\tau} = A$) realizes A *in exactly the same way* for *insubstantial* propositions. For this reason, how τ treats insubstantial propositions has no weight for the question whether τ is optimal.

3.2.5. FINDING OPTIMAL DYNAMIC ATTITUDES. We now give a characterization of the introspective propositional attitudes that strengthens our earlier Theorem 9, §1.7. According to that earlier result, the introspective propositional attitudes are just those propositional attitudes A such that there exists a dynamic attitude τ with $\overline{\tau} = A$. Now we can say something more:

THEOREM 40. *For a propositional attitude A, the following are equivalent:*

1. *A is introspective.*

2. *There exists a dynamic attitude τ that is optimal for A.*

PROOF. The right to left direction is obvious: if $\overline{\tau} = A$, then A is introspective, because fixed points of dynamic attitudes are introspective propositional attitudes by their definition. For the left to right direction, let A be a propositional attitude. Let f be a function that associates with each pair $(\mathcal{S}, P)$ such that $\mathcal{S}$ is a plausibility order, $P \subseteq W$, and $P \cap \mathrm{S} = P$ an element of $\mathrm{opt}_{\mathcal{S}} AP$. We now define a dynamic attitude τ as follows:

1. For any order $\mathcal{S}$ and proposition P: if $\mathcal{S} \vDash AP$, then $\mathcal{S}^{\tau P} := \mathcal{S}$.

2. For any order $\mathcal{S}$ and proposition P: if $\mathcal{S} \nvDash AP$ then

   — if P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} := \varnothing$,
   — if P is substantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} := f(\mathcal{S}, P \cap \mathrm{S})$.

By the construction, τ is optimal for A. To show this, the main point is to notice that, given a plausibility order $\mathcal{S}$ and a proposition P such that $\mathcal{S} \nvDash AP$ and P is substantial in $\mathcal{S}$, the function f picks a plausbility order $f(\mathcal{S}, P) \in \mathrm{opt}_{\mathcal{S}} AP$, which is to say that there exists no plausibility order $\mathcal{S}'$ such that $\mathcal{S}' \prec_{\mathcal{S}} f(\mathcal{S}, P)$ and $\mathcal{S}' \vDash AP$. ⊣

3.3. Non-Optimality

In this section, we provide examples of non-optimal dynamic attitudes.

3.3.1. Tests. Some typical examples of non-optimal dynamic attitudes are found among the tests. Consider $?K$, the test for irrevocable knowledge, which is (recall §1.7.4) defined by

$$\mathcal{S}^{?KP} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash KP, \\ \varnothing & \mathcal{S} \nvDash KP. \end{cases}$$

proposition 41. *$?K$ is not optimal.*

proof. The fixed point of $?K$ is irrevocable knowledge K. To see why $?K$ is not optimal for K (and thus not optimal), consider a plausibility order $\mathcal{S}$ such that $\mathrm{S} \cap P \neq \varnothing$ and $\mathrm{S} \cap \neg P \neq \varnothing$, i.e., S contains both P-worlds and non-P-worlds. By definition of $?K$: $\mathcal{S}^{?KP} = \varnothing$. On the other hand, $\mathcal{S}^{!P} \neq \varnothing$, hence $\mathrm{S}^{!P} \supset \mathrm{S}^{?KP}$, so $\mathcal{S}^{!P} <_{\mathcal{S}} \mathcal{S}^{?KP}$. Since $\bar{!} = K$, it follows that $?K$ is not optimal. ⊣

The change induced by the upgrade $?KP$ in an order $\mathcal{S}$ such that $\mathcal{S} \nvDash KP$ is thus "too drastic" to qualify for optimality. Deleting all worlds (as prescribed by $?K$) is not really needed; there are other ways of reaching the fixed point that preserve more structure. More precisely, whenever there are both P-worlds and non-P-worlds in $\mathcal{S}$, infallible trust $!$ does the job better.

3.3.2. Degrees of Trust vs. Spohn Revision. More examples of non-optimal dynamic attitudes arise by comparing two sets of dynamic attitudes that realize, as their fixed point, the degrees of belief introduced in §1.2.8.

For any natural number $n \geq 1$, the *n-Spohn revision* $\star_n$ is the dynamic attitude which associates with each given plausibility order $\mathcal{S}$ and proposition P the plausibility order $\mathcal{S}^{\star_n P}$ on the domain $\mathrm{S}^{\star_n P} = \mathrm{S}$, where for each world $w \in \mathrm{S}$, the Spohn ordinal $\kappa_{\mathcal{S}^{\star_n P}}(w)$ is given by

— if $P \cap \mathrm{S} = \varnothing$, then $\kappa_{\mathcal{S}^{\star_n P}}(w) = \kappa_{\mathcal{S}}(w)$.
— if $P \cap \mathrm{S} \neq \varnothing$, then
  — if $\kappa_{\mathcal{S}}(P) > n$, then
    — if $w \in P$, then $\kappa_{\mathcal{S}^{\star_n P}}(w) = \kappa_{\mathcal{S}}(w) - n$
    — if $w \notin P$, then $\kappa_{\mathcal{S}^{\star_n P}}(w) = \kappa_{\mathcal{S}}(w) + n$
  — if $\kappa_{\mathcal{S}}(P) \leq n$, then
    — if $w \in P$, then $\kappa_{\mathcal{S}^{\star_n P}}(w) = \kappa_{\mathcal{S}}(w) - (n - \kappa_{\mathcal{S}}(P))$

— if $w \notin P$, then $\kappa_{\mathcal{S}\star_n P}(w) = \kappa_{\mathcal{S}}(w) + (n - \kappa_{\mathcal{S}}(P))$

Spohn revision originates in Spohn's work on ranking functions.[9] The rough idea is to slide up *all* the P-worlds relative to the non-P-worlds. Figure 14.2 shows what is going on in a diagram, for the example of an upgrade $\star_n P$: the Spohn ordinal of all P-worlds is uniformly decremented until the best and the second-best P-worlds are best and second-best overall; at the same time, the Spohn ordinal of all non-P-worlds is uniformly incremented until the best non-P-worlds have Spohn degree 3.

Now contrast Spohn revision with the following dynamic attitude, which we define in the same format.

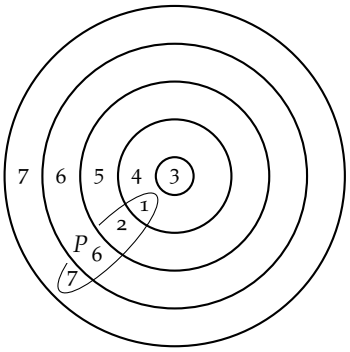
For any natural number $n \geq 1$, the *nth degree of trust* $\uparrow^n$ is the dynamic attitude which associates with each given plausibility order $\mathcal{S}$ and proposition P the plausibility order $\mathcal{S}^{\uparrow^n P}$ on the domain $\mathcal{S}^{\uparrow^n P} = \mathrm{S}$, where for any $w, v \in \mathrm{S}$:

— if $P \cap \mathrm{S} = \varnothing$, then $\kappa_{\mathcal{S}^{\uparrow^n P}}(w) = \kappa_{\mathcal{S}}(w)$.
— if $P \cap \mathrm{S} \neq \varnothing$, then
  — if $\kappa_{\mathcal{S}}(P) > n$, then
    — if $w \in P$ and $\kappa_{\mathcal{S}|_P}(w) \leq n$, then $\kappa_{\mathcal{S}^{\uparrow^n P}}(w) = \kappa_{\mathcal{S}}(w) - n$
    — if $w \notin P$ or $\kappa_{\mathcal{S}|_P}(w) > n$, then $\kappa_{\mathcal{S}^{\uparrow^n P}} = \kappa_{\mathcal{S}}(w) + n$
  — if $\kappa_{\mathcal{S}}(P) \leq n$, then
    — if $w \in P$ and $\kappa_{\mathcal{S}|_P}(w) \leq n$, then $\kappa_{\mathcal{S}^{\uparrow^n P}}(w) = \kappa_{\mathcal{S}}(w) - (n - \kappa_{\mathcal{S}}(P))$,
    — if $w \notin P$ or $\kappa_{\mathcal{S}|_P}(w) > n$, then $\kappa_{\mathcal{S}^{\uparrow^n P}} = \kappa_{\mathcal{S}}(w) + (n - \kappa_{\mathcal{S}}(P))$

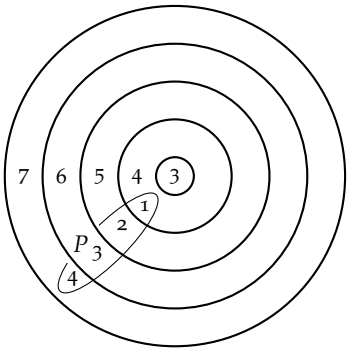
The family of attitudes $\{\uparrow^n\}_{n\in\omega}$ generalizes the idea of "promoting the best worlds" that underlies minimal trust: while $\uparrow P$ promotes the best P-worlds, making them better than everything else, $\uparrow^n P$ promotes the P-worlds that have Spohn ordinal up to n, making them better than everything else, while the order among the promoted worlds, and among the other, non-promoted worlds remains the same. Figure 14.1 shows what is going on in a diagram, for the example of an upgrade $\uparrow^2 P$, which promotes the best and the second-best P-worlds.

A glance at the two diagrams in 14.1 and 14.2 also reveals the difference between the two operations. On performing an upgrade $\uparrow^n P$, the P-worlds in $\mathcal{S}$ which have Spohn rank greater than n in $\mathcal{S}|_P$ are kept in place; but on performing an upgrade $\star_n P$, these same P-worlds also slide closer to the center of the associated system of spheres.

[9] Cf. Spohn (1988, 2009). Our definition looks much more complicated than the one given in Spohn (2009). It has the advantage, however, that it becomes apparent how P-worlds slide upwards, and non-P-worlds slide downwards, as an upgrade $\star_n P$ is applied.



(14.1)



(14.2)

FIGURE 14. Diagram (14.1) shows the result of performing an upgrade $\uparrow^2 P$ (where P is the proposition given by the ellipse) on the system of spheres given by the circles: the worlds with Spohn ordinal up to 2 are promoted towards the center, and otherwise the order remains the same. Diagram (14.2) shows the result of performing an upgrade $\star_2 P$ (where P is the proposition given by the ellipse) on the same system of spheres: the worlds with Spohn ordinal up to 2 are promoted towards the center, but in a rigid manner, so that the absolute distance between two given P-worlds is preserved.

Nevertheless, $\uparrow^n$ and $\star_n$ realize the same fixed point. We express this using the notion of a stricture (cf. §2.2.4):

PROPOSITION 42. *For any $n \in \omega$:*

— *The fixed point of the stricture of $\uparrow^n$ is the degree of belief B^n.*
— *The fixed poinf of the stricture of $\star_n$ is the degree of belief B^n.*

This observation makes $\uparrow^n$ and $\star_n$ interesting candidates for comparison in terms of our measure of similarity. From the perspective of our measure of similarity, however, Spohn revision preserves less structure than degrees of trust do:

PROPOSITION 43. *For any natural number n:*

1. *The degree of trust $\uparrow^n$ is optimal for B^n.*

2. *The Spohn revision $\star_n$ is not optimal for B^n.*

PROOF. The proof of item (1.) is similar to the proof of item (3.) of Proposition 37, so we omit it. For item (2.), we discuss the case where $n = 2$ as a representative sample case. Consider Figure 14. We start with an initial order, call it $\mathcal{S}$, given by the spheres drawn in the figure. We show that $\mathcal{S}^{\uparrow^2 P} <_{\mathcal{S}} \mathcal{S}^{\star_2 P}$ (notice

that the numbers in Figure 14.1 describe $\mathcal{S}^{\uparrow^2 P}$, while the numbers in Figure 14.1 describe $\mathcal{S}^{\star_2 P}$).

We first notice that $\mathrm{S}^{\uparrow^2 P} = \mathrm{S} = \mathrm{S}^{\star_2 P}$. To show our claim, it is thus sufficient to establish that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} \subset \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$, to which end we prove that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} \subseteq \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$, while it is not the case that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} = \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$.

To show that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} \subseteq \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$, let $w, v \in \mathrm{S}$. Observe:

— *If* $o_{\mathcal{S}}(w) \leq 3$ and $o_{\mathcal{S}}(v) \leq 3$, *then* $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_1 P}$ iff $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow P}$, and
— *else* $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^1 P}$.

These two observations are easy to check by inspecting Figure 14, and taken together they establish that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} \subseteq \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$.

To show that it is not the case that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} = \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$, take worlds $w, v \in \mathrm{S}$ such that $\kappa_{\mathcal{S}}(w) = \kappa_{\mathcal{S}}(v) = 4$, $w \in P \cap \mathrm{S}$, $v \in \neg P \cap \mathrm{S}$. Then $(v,w) \in \mathcal{S}$, $(v,w) \in \mathcal{S}^{\uparrow^2 P}$, while $(v,w) \notin \mathcal{S}^{\star_2 P}$. So it is not the case that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\star_2 P} = \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow^2 P}$. This yields the desired result: $\mathcal{S}^{\uparrow^2 P} \prec_{\mathcal{S}} \mathcal{S}^{\star_2 P}$, hence $\star_2$ is not optimal. ⊣

Thus, if one accepts the principle of minimal change, and, furthermore, accepts our measure of similarity introduced in the previous section, then Spohn revision $\star_n$ is not an optimal choice of a belief revision policy. We take this result mainly to indicate that the notion of optimality has real bite: not just any old dynamic attitude is optimal. The result should not be taken as an attempt to "prove" the inadequacy of Spohn revision. For it may well be argued that for Spohn's original framework (Spohn 1988, 2009), another measure of similarity is called for (which I will not attempt to spell out here, as this would take me too far from the main thread).

3.4. Canonicity

It may happen that a dynamic attitude τ is *uniquely optimal* in the sense that τ is the *only* optimal dynamic attitude whose fixed point is $\overline{\tau}$. If this is the case, we will call τ canonical.

3.4.1. Canonicity. Let τ be a dynamic attitude, and A a propositional attitude.

— τ is *canonical* if τ is optimal and for any attitude σ: if $\overline{\sigma} = \overline{\tau}$ and σ is optimal, then $\sigma = \tau$.

— τ is *canonical for* A if τ is canonical and $\overline{\tau} = A$.
— A is canonical if there exists a dynamic attitude that is canonical for A.

Our first observation is that there are canonical dynamic attitudes (§3.4.2); however, our second observation, not all propositional attitudes are canonical, or, equivalently: not all optimal dynamic attitudes are canonical (§3.4.3). The second observation raises a number of questions, which we begin to discuss towards the end of this section. They will keep us busy for the remainder of this chapter.

3.4.2. Some Canonical Dynamic Attitudes. Four of the five examples of dynamic attitudes mentioned in Proposition 37 are canonical for their fixed points:

proposition 44.

1. *Infallible trust* $!$ *is canonical.*

2. *Strong trust* $\Uparrow$ *is canonical.*

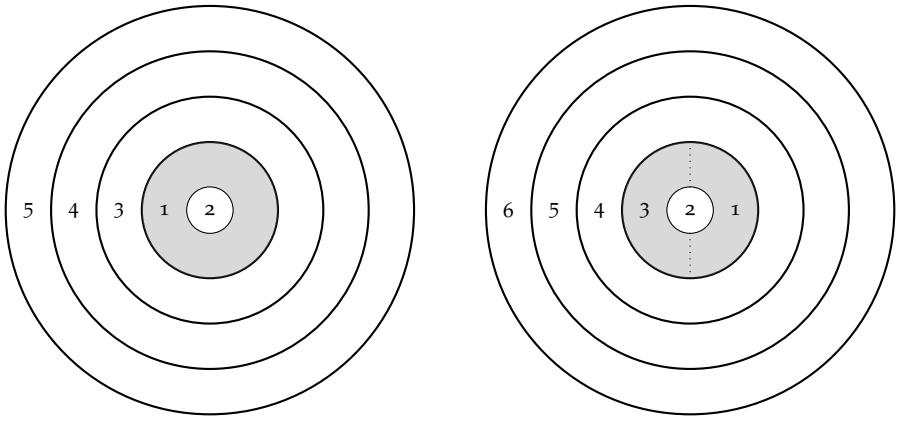
3. *Neutrality id is canonical.*

4. *Isolation* $\varnothing$ *is canonical.*

proof. These observations are consequences of Theorem 56 below. ⊣

Ahead of the proof of Theorem 56, the intuition why these dynamic attitudes are canonical is clear: for each of these dynamic attitudes, there is *really only one thing one can do* to reach the corresponding fixed point in a minimal way: "delete all P-worlds, keep all else the same" to realize K; and "make all P-worlds better than all non-P-worlds, keep all else the same" to realize $Sb \vee K^{\neg}$; and "keep everything the same" to realize $\top$; and "delete everything" to realize $\bot$.

If one accepts the principle of minimal change (and our formalization of it), then the dynamic attitudes considered in the above proposition are thus the *only reasonable choice* for dynamic attitudes aiming at their respective fixed points. And if one does not subscribe to the principle, the result still provides an interesting characterization of the respective dynamic attitudes and their fixed points.



(15.1) (15.2)

FIGURE 15. Two ways of ensuring that the most plausible worlds are gray. In Diagram 15.1, *all* gray worlds are promoted towards the center, while in Diagram 15.2, only the gray worlds towards the right of the dotted line are promoted towards the center.

3.4.3. SOME NON-CANONICAL DYNAMIC ATTITUDES. Our next observation is that there are optimal dynamic attitudes that are not canonical. We discuss three examples of this phenomenon. In each case, the proof relies on a characterization of canonicity for propositional attitudes that we only formally establish in §3.6 below. According to this characterization (Corollary 55), a propositional attitude A is canonical iff for every plausibility order $\mathcal{S}$ and $\mathcal{S}$-substantial P, there exists an order $\mathcal{S}' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \vDash AP$ and for any order $\mathcal{S}'' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \neq \mathcal{S}''$: if $\mathcal{S}'' \vDash AP$, then $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$. In view of this characterization, the following proof strategy, used in the proof of the proposition below, is sound: to establish non-canonicity of a propositional attitude A, find a plausibility order $\mathcal{S}$, and an $\mathcal{S}$-substantial proposition P, and find orders $\mathcal{S}', \mathcal{S}'' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \vDash AP$, $\mathcal{S}'' \vDash AP$, while $\mathcal{S}'$ and $\mathcal{S}''$ are *incomparable* in the relation $\prec_{\mathcal{S}}$, that is: neither $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$, nor $\mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}'$. Applying the result cited above, one may then conclude that A is not canonical. And from this, it follows that any dynamic attitude whose fixed point is A is not canonical either.

PROPOSITION 45.

1. *Simple belief B is not canonical.*

2. *Refinedness R is not canonical.*[10]

[10]Recall the definition of refinedness R from §2.5.1: $\mathcal{S} \vDash RP$ iff $\forall w, v \in \mathrm{S} : w \in P, v \notin P \Rightarrow w \not\in$

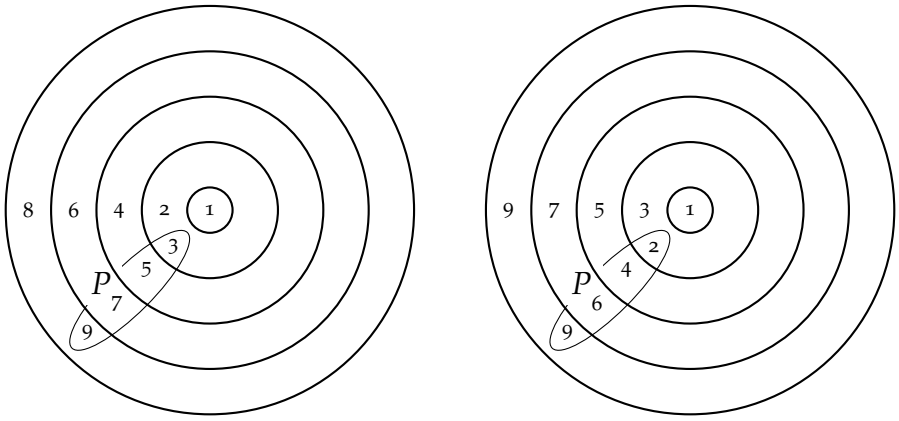


FIGURE 16. Two ways of converting the input order into a plausibility order satisfying RP (where R is the propositional attitude *refinedness*, with RP satisfied in a plausibility order iff there are no ties between P-worlds and non-P-worlds). In Figure (16.1), ties between P-worlds and non-P-worlds are resolved in favour of the non-P-worlds, while in Figure (16.2), they are resolved in favour of the P-worlds.

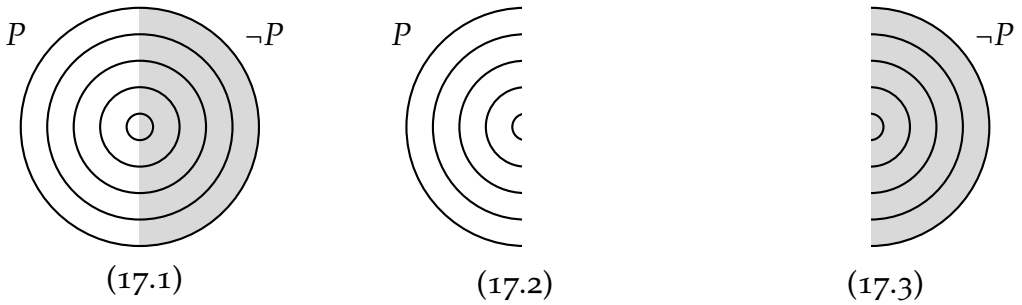


FIGURE 17. Diagram 17.2 and Diagram 17.3 depict two ways of converting the input order shown in Diagram 17.1 into a plausibility order satisfying $KP \vee K^{\neg}P$. In Diagram (17.2), this is achieved by deleting all P-worlds, while in Diagram (17.2) it is achieved by deleting all non-P-worlds.

3. *The disjunction of knowledge and opposite knowledge $K \vee K^{\neg}$ is not canonical.*

PROOF. For each item, we supply an order $\mathcal{S}$, a proposition P substantial in $\mathcal{S}$, and orders $\mathcal{S}', \mathcal{S}'' \in O_{\mathcal{S}}$ such that neither $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$ nor $\mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}'$. For the first item, cf. Figure 15; for the second item, cf. Figure 16; for the third item,

v. The attitude R should not be confused with refined belief Rb, as defined in §1.2.8.

cf. Figure 17. For each item, the claim then follows from Corollary 55. ⊣

To my mind, these three canonicity failures fall into two distinct categories. The fact that R and $K \vee K^{\neg}$ are not canonical seems to be related to intrinsic characteristics of these propositional attitudes: I cannot think of a reasonable formalization of the concept of minimal change that would give rise to a unique way of transforming given plausibility orders to orders satisfying RP, or $KP \vee K\neg P$. As for the latter, for example: the choice between deleting the P-worlds and deleting the non-P-worlds is, intuitively speaking, *completely arbitrary*, and no amount of formal work should be expected to do anything about this. The case of simple belief B, on the other hand, is different. One easily gets the feeling that there is something quite special, and, indeed, unique, about the dynamic attitude $\uparrow^+$. Indeed, many authors have seen Boutilier's minimal revision (which is simply a slight variant of $\uparrow^+$, corresponding to our $\uparrow$) as *the* embodiment of minimality of belief change. However, our theory says otherwise: $\uparrow^+$ is optimal for belief, but, as a consequence of the above proposition, not canonical. This raises the question whether our formal apparatus should be adjusted in some way so as to capture in what sense $\uparrow^+$ is unique.

Taking a step back, the results of this section raise two questions:

1. Given that not all propositional attitudes are canonical—which ones are?

2. How can the theory be amended to more closely match our intuitions about canonicity?

The purpose of §3.5 and §3.6 is to make progress towards answering the first question, while the final section of this chapter, §3.7, addresses the second one, studying it concretely for the case of simple belief.

3.5. Characterizing Optimality and Canonicity

The definitions of optimality and canonicity are somewhat awkward to work with: to check whether a dynamic attitude τ is optimal, we need to compare it to arbitrary *other* dynamic attitudes with the same fixed point; to check whether τ is canonical, we need, first, to check that it is optimal, and, second, to check whether it is *unique* in that respect, which amounts to showing that any dynamic attitude that is optimal for $\overline{\tau}$ is actually τ itself. It would be nice if one did not have to do this on a case by case basis.

As a first step, we would like to characterize optimality and canonicity in a way that only depends on the notion of similarity. Intuitively, it is not

hard to see how such a characterization should look. Recall that a dynamic attitude τ is optimal if there is no other dynamic attitude σ such that, given some proposition P, there exists a plausibility order $\mathcal{S}$ such that $\mathcal{S}^{\sigma P}$ is more similar to $\mathcal{S}$ than $\mathcal{S}^{\tau P}$ (cf.§3.2.1). What we would like to see is that this is the case iff, given some plausibility order $\mathcal{S}$, and some proposition P, applying τP picks in general a closest order $\mathcal{S}'$ satisfying $\mathcal{S}' \vDash \overline{\tau} P$; and canonical if it always picks *the* (unique) closest order $\mathcal{S}'$ such that $\mathcal{S}' \vDash \overline{\tau} P$.

Roughly speaking, these are indeed the results we obtain, and the purpose of the present section is to demonstrate this. Start with optimality. Proposition 47 below provides a characterization of the optimal dynamic attitudes in terms of the closest orders satisfying their fixed point. We first prove an auxiliary observation.

LEMMA 46. *Let τ be a dynamic attitude. Let $\mathcal{S}_1, \mathcal{S}_2$ be plausibility orders such that $S_2 \subseteq S_1$. Suppose that P is $\mathcal{S}_1$-substantial. Assume that $\mathcal{S}_1 \nvDash \overline{\tau} P$, $\mathcal{S}_2 \vDash \overline{\tau} P$. Then there exists a dynamic attitude σ such that $\overline{\sigma} = \overline{\tau}$ and $\mathcal{S}_1^{\sigma P} = \mathcal{S}_2$.*

PROOF. Define the dynamic attitude σ as follows:

$$\mathcal{S}^{\sigma Q} := \begin{cases} \mathcal{S}_2 & \mathcal{S} = \mathcal{S}_1, Q \cap S = P \cap S, \\ \mathcal{S}^{\tau Q} & \text{otherwise.} \end{cases}$$

Clearly, σ meets our requirements. ⊣

PROPOSITION 47. *Let τ be a dynamic attitude. The following are equivalent:*

1. *τ is optimal.*

2. *For any $\mathcal{S}$ and $\mathcal{S}$-substantial P: $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$.*

PROOF. From (1.) to (2.), we argue by contraposition. Suppose that there exists an order $\mathcal{S}$ and an $\mathcal{S}$-substantial proposition P such that $\mathcal{S}^{\tau P} \notin \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$. Choose an element $\mathcal{S}' \in \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ such that $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}^{\tau P}$ (guaranteed to exist by Lemma 38: $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ is non-empty). We observe that $\mathcal{S} \neq \mathcal{S}'$, hence we know that $\mathcal{S} \nvDash \overline{\tau} P$ and $\mathcal{S}' \vDash \overline{\tau} P$. So we may apply Lemma 46 to conclude that there exists a dynamic attitude σ such that $\mathcal{S}^{\sigma P} = \mathcal{S}'$. Since $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\tau P}$, it follows that τ is not optimal.

From (2.) to (1.), we argue again by contraposition. Suppose that τ is not optimal. Then there exists an order $\mathcal{S}$, a proposition P, and an attitude σ such that $\overline{\sigma} = \overline{\tau}$ and $\mathcal{S}^{\sigma P} \prec_{\mathcal{S}} \mathcal{S}^{\tau P}$. By Lemma 39, P is $\mathcal{S}$-substantial (for otherwise, $\mathcal{S}^{\sigma P} = \mathcal{S}^{\tau P}$, contradiction). Since $\mathcal{S}^{\sigma P} \vDash \overline{\tau} P$, it follows that $\mathcal{S}^{\tau P} \notin \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$, the desired result. ⊣

So optimal dynamic attitudes are, indeed, those that generally pick some closest output order realizing their fixed point, given a propositional input that is substantial in the input order. We notice that this implies that optimality is invariant under strictures:

COROLLARY 48. *For any dynamic attitude τ: τ is optimal iff τ^+ is optimal.*

PROOF. Let τ be a dynamic attitude. By Theorem 47, τ is optimal iff for any plausibility order $\mathcal{S}$ and proposition P substantial in $\mathcal{S}$: $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$. But for any such $\mathcal{S}$ and P: $\mathcal{S}^{\tau^+ P} = \mathcal{S}^{\tau P}$ by definition of stricture (cf. §2.6.4, the last item in the list). So τ is optimal iff τ^+ is optimal. ⊣

Next, we extend our analysis to canonicity.

LEMMA 49. *Let τ be an optimal dynamic attitude. Let $\mathcal{S}_1, \mathcal{S}_2$ be plausibility orders such that $\mathrm{S}_2 \subseteq \mathrm{S}_1$. Suppose that P is $\mathcal{S}_1$-substantial. Assume that $\mathcal{S}_1 \not\models \overline{\tau} P$, $\mathcal{S}_2 \models \overline{\tau} P$. Assume that $\mathcal{S}_2 \in \mathrm{opt}_{\mathcal{S}} \overline{\tau} P$. Then there exists an optimal dynamic attitude σ such that $\overline{\sigma} = \overline{\tau}$ and $\mathcal{S}_1^{\sigma P} = \mathcal{S}_2$.*

PROOF. We define σ as in the proof of Lemma 46:

$$\mathcal{S}^{\sigma Q} := \begin{cases} \mathcal{S}_2 & \mathcal{S} = \mathcal{S}_1, Q \cap \mathrm{S} = P \cap \mathrm{S}, \\ \mathcal{S}^{\tau Q} & \text{otherwise.} \end{cases}$$

Clearly, σ meets our requirements. ⊣

PROPOSITION 50. *If τ is canonical, then for any $\mathcal{S}$ and $\mathcal{S}$-substantial P: $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ is a singleton set.*

PROOF. We show the contrapositive. Suppose that there exists an order $\mathcal{S}$, a proposition P such that P is substantial in $\mathcal{S}$, and suppose that $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ is not a singleton. Since, generally, $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P_\star \neq \varnothing$ (Lemma 38), this means that $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ has at least two elements.

Choose an element of $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ such that $\mathcal{S}' \neq \mathcal{S}^{\tau P}$ (guaranteed to exist by the previous step). By Lemma 49, there exists an optimal dynamic attitude σ such that $\overline{\sigma} = \overline{\tau}$ and $\mathcal{S}^{\sigma P} = \mathcal{S}'$. But σ is distinct from τ, since $\mathcal{S}^{\sigma P} \neq \mathcal{S}^{\tau P}$. So τ is not canonical. This shows our claim. ⊣

PROPOSITION 51. *Suppose that for any $\mathcal{S}$ and $\mathcal{S}$-substantial P: $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ is a singleton. If τ is optimal, then τ is canonical.*

PROOF. Assume that for any order $\mathcal{S}$ and proposition P such that P is substantial in $\mathcal{S}$: $\mathrm{opt}_{\mathcal{S}} \overline{\tau} P$ is a singleton. Suppose that τ is optimal. It follows

from Proposition 47 that for any order $\mathcal{S}$ and proposition P substantial in $\mathcal{S}$: $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P$.

Suppose now that some dynamic attitude σ is optimal for $\overline{\tau}$. Let $\mathcal{S}$ be a plausibility order and P a proposition. If P is insubstantial in $\mathcal{S}$, by Lemma 39, $\mathcal{S}^{\sigma P} = \mathcal{S}^{\tau P}$. If P is substantial in $\mathcal{S}$, by Proposition 47, using the fact that $\mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P$ is a singleton, $\mathcal{S}^{\sigma P} = \mathcal{S}^{\tau P}$. So $\sigma = \tau$. By definition, τ is canonical, which completes the proof. ⊣

We now show that canonical dynamic attitudes are those that in general pick *the unique closest order* to realize their fixed point, given a substantial proposition as an input.

THEOREM 52. *Let τ be a dynamic attitude. The following are equivalent:*

1. *τ is canonical.*

2. *For any $\mathcal{S}$ and $\mathcal{S}$-substantial P: $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P$ and $|\mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P| = 1$.*

PROOF. For the direction from (1.) to (2.), suppose that τ is canonical (and hence optimal). Let $\mathcal{S}$ be a plausibility order, suppose that P is substantial in $\mathcal{S}$. By Proposition 50, the set $\mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P$ is a singleton (one part of our claim). Since τ is optimal, by Proposition 47, $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}}\,\overline{\tau}P$ (the second part of our claim). This shows the direction form (1.) to (2.).

For the converse direction, assume that the condition stated in (2.) holds. By Proposition 47, τ is optimal. By Proposition 51, τ is canonical. This shows the converse direction, and completes the proof. ⊣

The result says that a dynamic attitude τ is canonical iff there is a *unique way of realizing its fixed point $\overline{\tau}$ in a minimal way* for propositions that are substantial in given orders. More loosely speaking: τ is canonical iff the principle of minimal change fully determines its behaviour. And this is, of course, what we would like our notion of canonicity to amount to.

3.6. Canonical Propositional Attitudes

Recall Proposition 45: certain propositional attitudes are not canonical. This observation leads to the question already suggested above: what characterizes propositional attitudes that *are* canonical? In this section, we make progress towards answering this question. A first answer to the question will follow from the work of the previous section: Theorem 54 characterizes the canonical propositional attitudes A as those attitudes which allow us to find, for any

given order $\mathcal{S}$ and $\mathcal{S}$-substantial proposition P, a unique closest order satisfying AP. This result is, however, not as illuminating as one would like it to be. We will thus continue working towards Theorem 56, the main result of this section, which gives sufficient criteria for canonicity in terms of two simple preservation properties.

We start our work with the following lemma.

LEMMA 53. *For any order $\mathcal{S}$, proposition P and introspective propositional attitude A:* $\mathrm{opt}_{\mathcal{S}} AP = \mathrm{opt}_{\mathcal{S}} A(P \cap \mathrm{S})$.

PROOF. Let $\Theta = \{\mathcal{S}' \mid \mathcal{S} \twoheadrightarrow \mathcal{S}', \mathcal{S}' \vDash AP\}$. Notice that $\Theta = \{\mathcal{S}' \mid \mathcal{S} \twoheadrightarrow \mathcal{S}', \mathcal{S}' \vDash A(P \cap \mathrm{S}')\}$ by conservativity of A (cf. §1.2.2). Observing that $(P \cap \mathrm{S}') = (P \cap \mathrm{S}) \cap \mathrm{S}'$, it follows that $\Theta = \{\mathcal{S}' \mid \mathcal{S} \twoheadrightarrow \mathcal{S}', \mathcal{S}' \vDash A((P \cap \mathrm{S}) \cap \mathrm{S}')\}$. But then, again by conservativity of A, it follows that $\Theta = \{\mathcal{S}' \mid \mathcal{S} \twoheadrightarrow \mathcal{S}', \mathcal{S}' \vDash A(P \cap \mathrm{S})\}$. So $\mathrm{opt}_{\mathcal{S}} AP = \{\mathcal{S}' \in \Theta \mid \neg\exists \mathcal{S}'' \in \Theta : \mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}'\} = \mathrm{opt}_{\mathcal{S}} A(P \cap \mathrm{S})$. ⊣

We now obtain the following corollary from Theorem 52:

COROLLARY 54. *Let A be an introspective propositional attitude. The following are equivalent:*

1. *A is canonical.*

2. *For any $\mathcal{S}$ and $\mathcal{S}$-substantial P:* $|\mathrm{opt}_{\mathcal{S}} AP| = 1$.

PROOF. From (1.) to (2.), suppose that A is canonical. Then there exists a canonical dynamic attitude τ such that $\overline{\tau} = A$. By Theorem 52, $|\mathrm{opt}_{\mathcal{S}} AP| = 1$ for any $\mathcal{S}$ and $\mathcal{S}$-substantial P.

From (2.) to (1.), suppose that for any $\mathcal{S}$ and $\mathcal{S}$-substantial P: $|\mathrm{opt}_{\mathcal{S}} AP| = 1$. Let $\mathcal{S}$ be a plausibility order. Define a dynamic attitude τ as follows:

— If P is insubstantial in $\mathcal{S}$, put

$$\mathcal{S}^{\tau P} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash AP, \\ \varnothing & \mathcal{S} \nvDash AP. \end{cases}$$

— If P is substantial in $\mathcal{S}$, put

$$\mathcal{S}^{\tau P} := \mathcal{S}', \text{ where } \mathcal{S}' \text{ is the unique element of the singleton } \mathrm{opt}_{\mathcal{S}} AP.$$

We have to show that τ is indeed a dynamic attitude. By Proposition 6, we have to show that τ satisfies (1) strong informativity, (2) idempotence and (3) dynamic conservativity.

1. If $P \in \{W, \varnothing\}$, then P is insubstantial in $\mathcal{S}$, hence $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$ by definition of τ. So τ satisfies strong informativity.

2. If P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} \models AP$ by definition of τ. But if $\mathcal{S}^{\tau P} \models AP$, then $(\mathcal{S}^{\tau P})^{\tau P} = \mathcal{S}^{\tau P}$, again by definition of τ. If, on the other hand, P is substantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} \in \mathrm{opt}_{\mathcal{S}} AP$ by definition of τ. But for any $\mathcal{S}' \in \mathrm{opt}_{\mathcal{S}} AP$: $\mathrm{opt}_{\mathcal{S}'} AP = \{\mathcal{S}'\}$. Hence $(\mathcal{S}^{\tau P})^{\tau P} = \mathcal{S}^{\tau P}$. So τ is idempotent.

3. Notice that $\mathcal{S} \models AP$ iff $\mathcal{S} \models A(P \cap \mathrm{S})$ by conservativity of A. By definition of τ, it follows that $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau(P \cap \mathrm{S})}$ for any order $\mathcal{S}$ and proposition P that is insubstantial in $\mathcal{S}$. Now suppose that P is substantial in $\mathcal{S}$. By Lemma 53, $\mathrm{opt}_{\mathcal{S}} AP = \mathrm{opt}_{\mathcal{S}} A(P \cap \mathrm{S})$. It follows from the initial assumption that $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau(P \cap \mathrm{S})}$, using the definition of τ. So τ satisfies dynamic conservativity.

We conclude that τ is a dynamic attitude. But $\overline{\tau} = A$ by definition of τ. By Theorem 52, τ is canonical for A. Hence A is canonical. ⊣

The following reformulation has the advantage that it can be applied more directly to conclude that a propositional attitude is *not* canonical (as we have already used it in Proposition 45 above):

COROLLARY 55. *Let A be an introspective propositional attitude. The following are equivalent:*

1. *A is canonical.*

2. *For every plausibility order $\mathcal{S}$ and $\mathcal{S}$-substantial P, there exists an order $\mathcal{S}' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \models AP$ and for any order $\mathcal{S}'' \in O_{\mathcal{S}}$ such that $\mathcal{S}' \neq \mathcal{S}''$: if $\mathcal{S}'' \models AP$, then $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$.*

PROOF. For the direction from (2.) to (1.), observe that from (2.), we easily infer that for any order $\mathcal{S}$ and $\mathcal{S}$-substantial P: $|\mathrm{opt}_{\mathcal{S}} AP| = 1$, and the claim follows using Corollary 54.

For the direction from (1.) to (2.), suppose that A is canonical, and thus, by Corollary 54: for any order $\mathcal{S}$ and $\mathcal{S}$-substantial P: $|\mathrm{opt}_{\mathcal{S}} AP| = 1$. Take a plausibility order $\mathcal{S}$ and an $\mathcal{S}$-substantial proposition P, and let $\mathcal{S}'$ be the unique element of the singleton set $|\mathrm{opt}_{\mathcal{S}} AP| = 1$. Take any order $\mathcal{S}'' \in O_{\mathcal{S}}$ distinct from $\mathcal{S}'$ satisfying $\mathcal{S}'' \models AP$; we claim that $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$. Suppose otherwise. Then there exists an order $\mathcal{S}''' \in \mathrm{opt}_{\mathcal{S}} AP$ such that $\mathcal{S}''' \preceq_{\mathcal{S}} \mathcal{S}''$. Clearly, $\mathcal{S}''' \neq \mathcal{S}'$ (for otherwise, $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$, contradiction). It follows that $|\mathrm{opt}_{\mathcal{S}} AP| \neq 1$. This is a contradiction, so $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}''$, after all. This shows that (2.) holds. ⊣

3.6.1. Refinements and Joint Embeddings. As already pointed out above, both Corollary 54 and Corollary 55 leave something to be desired. A more insightful characterization would break down the notion of canonicity into a number of (jointly) necessary and (jointly) sufficient conditions that are conceptually illuminating and easy to check. We answer "one half" of this query by providing two conditions that are jointly sufficient for canonicity. Here are the conditions:

— Let $\mathcal{S}$, $\mathcal{S}'$ be plausibility orders. $\mathcal{S}'$ is a *refinement* of $\mathcal{S}$ (or: $\mathcal{S}'$ *refines* $\mathcal{S}$) if $S = S'$ and $\mathcal{S}' \cup \{(w,v) \mid w \approx_{\mathcal{S}} v\} = \mathcal{S}$. We say that A is *preserved under refinements* iff for any plausibility orders $\mathcal{S},\mathcal{S}'$ and proposition P: if $\mathcal{S} \vDash AP$ and $\mathcal{S}'$ is a refinement of $\mathcal{S}$, then $\mathcal{S}' \vDash AP$.
— Let $\mathcal{S}$, $\mathcal{S}'$ be a plausibility order. $\mathcal{S}'$ *embeds into* $\mathcal{S}$ if $\mathcal{S}' \subseteq \mathcal{S}$. We say that A *admits joint embeddings* iff for any plausibility orders $\mathcal{S}_1$ and $\mathcal{S}_2$ and proposition P: if $\mathcal{S}_1 \vDash AP$ and $\mathcal{S}_2 \vDash AP$, then there exists an order $\mathcal{S}'$ such that $\mathcal{S}' \vDash AP$ and $\mathcal{S}_1$ and $\mathcal{S}_2$ each embed into $\mathcal{S}'$, and $S' = S_1 \cup S_2$.

Our main result is the following:

theorem 56. *If A is preserved under refinements and admits joint embeddings, then A is canonical.*

proof. The result is a direct consequence of Proposition 57 and Proposition 58 below. We state and prove them in turn.

proposition 57. *Suppose that for all plausibility orders $\mathcal{S}$ and propositions P substantial in $\mathcal{S}$, the following holds:*

> *If $\mathcal{S} \twoheadrightarrow \mathcal{S}_1$, $\mathcal{S} \twoheadrightarrow \mathcal{S}_2$, $\mathcal{S}_1 \vDash AP$, and $\mathcal{S}_2 \vDash AP$, then $\exists \mathcal{S}'$: $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_1$, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_2$, and $\mathcal{S}' \vDash AP$.*

Then A is canonical.

proof. Suppose that for all $\mathcal{S}$, for all substantial P, the following holds: If $\mathcal{S}_1 \vDash AP$, $\mathcal{S}_2 \vDash AP$, and $\mathcal{S} \twoheadrightarrow \mathcal{S}_1$, $\mathcal{S} \twoheadrightarrow \mathcal{S}_2$, then $\exists \mathcal{S}'$: $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_1$, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_2$, and $\mathcal{S}' \vDash AP$.

Our aim is to show that A is canonical. For this purpose, let $\mathcal{S}$ be a plausibility order, and let P be a proposition substantial in $\mathcal{S}$. By Theorem 54, it is sufficient to establish that $\mathrm{opt}_{\mathcal{S}} AP$ is a singleton set.

Proceeding to prove this, take two orders $\mathcal{S}_1, \mathcal{S}_2 \in \mathrm{opt}_{\mathcal{S}} AP$. We show that $\mathcal{S}_1 = \mathcal{S}_2$, which proves our claim.

By the assumption, there exists $\mathcal{S}'$ such that $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_1$, $\mathcal{S}' \leq_{\mathcal{S}} \mathcal{S}_2$, $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, $\mathcal{S}' \vDash AP$.

Since $\mathcal{S}_1, \mathcal{S}_2 \in \text{opt}_{\mathcal{S}} AP$, it is not the case that $\mathcal{S}' <_{\mathcal{S}} \mathcal{S}_1$, and it is not the case that $\mathcal{S}' <_{\mathcal{S}} \mathcal{S}_2$. Hence $\mathcal{S}' = \mathcal{S}_1$, $\mathcal{S}' = \mathcal{S}_2$, and thus $\mathcal{S}_1 = \mathcal{S}_2$, which is just what we need. ⊣

To prove the claim of Theorem 56, it is then sufficient to show the following:

PROPOSITION 58. *Suppose that A admits joint embeddings and A is preserved under refinements. Then the following holds for all plausbility orders $\mathcal{S}$ and propositions P substantial in $\mathcal{S}$:*

> *If $\mathcal{S} \twoheadrightarrow \mathcal{S}_1$, $\mathcal{S} \twoheadrightarrow \mathcal{S}_2$, $\mathcal{S}_1 \vDash AP$, and $\mathcal{S}_2 \vDash AP$, then $\exists \mathcal{S}'$: $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_1$, $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_2$, and $\mathcal{S}' \vDash AP$.*

PROOF. Suppose that A admits joint embeddings and A is preserved under refinements.

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be plausibility orders such that $\mathcal{S}_1 \vDash AP$, $\mathcal{S}_2 \vDash AP$, $\mathcal{S} \twoheadrightarrow \mathcal{S}_1$, $\mathcal{S} \twoheadrightarrow \mathcal{S}_2$, with P substantial in $\mathcal{S}$.

We need to find a plausibility order $\mathcal{S}'$ such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_1$, $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_2$, and $\mathcal{S}' \vDash AP$.

Our first observation is that, by the fact that A admits joint embeddings, there exists a plausibility order $\mathcal{S}_\#$ such that $\mathcal{S}_\# \supseteq \mathcal{S}_1$, $\mathcal{S}_\# \supseteq \mathcal{S}_2$, $\mathrm{S}_\# = \mathrm{S}_1 \cup \mathrm{S}_2$ and $\mathcal{S}_\# \vDash AP$. We observe that $\mathcal{S} \twoheadrightarrow \mathcal{S}_\#$.

We now prove an auxiliary observation. For any plausibility orders $\mathcal{S}, \mathcal{T}$ such that $\mathcal{S} \twoheadrightarrow \mathcal{T}$, let

$$agree^+_{\mathcal{S}}\mathcal{T} := \{(w,v) \in \mathrm{S} \times \mathrm{S} \mid (w,v) \in \mathcal{T}, (w,v) \in \mathcal{S}\},$$

$$agree^-_{\mathcal{S}}\mathcal{T} := \{(w,v) \in \mathrm{S} \times \mathrm{S} \mid (w,v) \notin \mathcal{T}, (w,v) \notin \mathcal{S}\}.$$

We observe that, in general, $agree_{\mathcal{S}}\mathcal{T} = agree^+_{\mathcal{S}}\mathcal{T} \cup agree^-_{\mathcal{S}}\mathcal{T}$.

Returning to our orders $\mathcal{S}_\#$, $\mathcal{S}_1$ and $\mathcal{S}_2$, we prove, for later use, the following claim:

CLAIM.

1. $agree^+_{\mathcal{S}}\mathcal{S}_\# \supseteq agree^+_{\mathcal{S}}\mathcal{S}_1 \cup agree^+_{\mathcal{S}}\mathcal{S}_2$.

2. *If $(w,v) \in agree^-_{\mathcal{S}}\mathcal{S}_1 \cup agree^-_{\mathcal{S}}\mathcal{S}_2$ and $(w,v) \notin agree^-_{\mathcal{S}}\mathcal{S}_\#$, then $w \approx_{\mathcal{S}_\#} v$.*

PROOF. For item (1.), suppose that $(w,v) \in agree^+_{\mathcal{S}}\mathcal{S}_1$. Then $(w,v) \in \mathcal{S}_1$, so $(w,v) \in \mathcal{S}_\#$, and also $(w,v) \in \mathcal{S}$, so $(w,v) \in agree^+_{\mathcal{S}}\mathcal{S}_\#$. Under the supposition that $(w,v) \in agree^+_{\mathcal{S}}\mathcal{S}_2$, make the analogous case. The claim follows.

For item (2.), suppose that $(w,v) \in agree^-_{\mathcal{S}}\mathcal{S}_1$, $(w,v) \notin agree^-_{\mathcal{S}}\mathcal{S}_\#$. Since $(w,v) \in agree^-_{\mathcal{S}}\mathcal{S}_1$, it follows that $(w,v) \notin \mathcal{S}_1$, $(w,v) \notin \mathcal{S}$. Since $(w,v) \notin agree^-_{\mathcal{S}}\mathcal{S}_\#$, it follows that $(w,v) \in \mathcal{S}_\#$. But since $(w,v) \notin \mathcal{S}_1$, it follows that $(v,w) \in \mathcal{S}_1$, hence $(v,w) \in \mathcal{S}_\#$. So $w \approx_{\mathcal{S}_\#} v$, the desired result. ⊣

We now return to the main thread, and use the plausibility order $\mathcal{S}_\#$ defined above to find the desired plausibility order $\mathcal{S}'$.

Define $\mathcal{S}' = (S', \leq_{\mathcal{S}'})$ by means of setting $S' := S_1 \cup S_2$, and requiring, for all $w, v \in S'$:

— if $w \not\approx_{\mathcal{S}_\#} v$, then $w \leq_{\mathcal{S}'} v$ iff $w \leq_{\mathcal{S}_\#} v$.
— if $w \approx_{\mathcal{S}_\#} v$, then $w \leq_{\mathcal{S}'} v$ iff $w \leq_{\mathcal{S}} v$.

We claim that $\mathcal{S}'$ is a plausibility order. Reflexivity and connectedness are immediate by the fact that $\mathcal{S}$ and $\mathcal{S}_\#$ are both reflexive and connected. To show that $\mathcal{S}'$ is transitive, we argue as follows. Suppose that $(w, v), (v, x) \in \mathcal{S}'$. We distinguish four possible cases, which we discuss in turn:

— *Case 1:* Suppose that $w \not\approx_{\mathcal{S}_\#} v$ and $v \not\approx_{\mathcal{S}_\#} x$. By the assumption, $w <_{\mathcal{S}_\#} v <_{\mathcal{S}_\#} x$, hence $w <_{\mathcal{S}_\#} x$. So $(w, x) \in \mathcal{S}'$ by definition of $\mathcal{S}'$.
— *Case 2:* Suppose that $w \approx_{\mathcal{S}_\#} v$ and $v \approx_{\mathcal{S}_\#} x$, i.e., $w \approx_{\mathcal{S}_\#} v \approx_{\mathcal{S}_\#} x$, and thus $w \approx_{\mathcal{S}_\#} x$. So $(w, x), (x, v) \in \mathcal{S}$ by the assumption, $(w, v) \in \mathcal{S}$ by transitivity of $\mathcal{S}$. So $(w, x) \in \mathcal{S}'$ by definition of $\mathcal{S}'$.
— *Case 3:* Suppose that $w \not\approx_{\mathcal{S}_\#} v$ and $v \approx_{\mathcal{S}_\#} x$. By the assumption, $w <_{\mathcal{S}_\#} v \approx_{\mathcal{S}_\#} x$, hence $w <_{\mathcal{S}_\#} x$. So $(w, x) \in \mathcal{S}'$ by definition of $\mathcal{S}'$.
— *Case 4:* Suppose that $w \approx_{\mathcal{S}_\#} v$ and $v \not\approx_{\mathcal{S}_\#} x$. Then, using the assumption, $w \approx_{\mathcal{S}_\#} v <_{\mathcal{S}_\#} x$, hence $w <_{\mathcal{S}_\#} x$. So $(w, x) \in \mathcal{S}'$ by definition of $\mathcal{S}'$.

In each case, $(w, x) \in \mathcal{S}'$, hence $\mathcal{S}'$ is transitive, and thus indeed a plausibility order.

We now argue that $\mathcal{S}'$ has all the properties we want. First, notice that $\mathcal{S}'$ is a refinement of $\mathcal{S}_\#$, hence, by the assumption that A is preserved under refinements, and recalling that $\mathcal{S}_\# \models AP$, we conclude that $\mathcal{S}' \models AP$. Furthermore, $\mathcal{S} \twoheadrightarrow \mathcal{S}'$. It remains to be shown that $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_1$ and $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_2$.

To prove this, we show that $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}_1 \cup agree_{\mathcal{S}}\mathcal{S}_2$. This entails that $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_1$ and $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_2$, for the following reason: take $\mathcal{S}_i$ with $i \in \{1, 2\}$. We know that $S' \supseteq S_i$. If $S' \supset S$, then $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}_i$, hence $\mathcal{S}' \preceq_{\mathcal{S}} \mathcal{S}_i$. If, on the other hand, $S' = S$, it is sufficient to show that $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}_i$ to be able to conclude that $\mathcal{S}' \prec_{\mathcal{S}} \mathcal{S}_i$. Overall, this implies that if we are able to show that $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}_1 \cup agree_{\mathcal{S}}\mathcal{S}_2$, we are done.

This, then, is our final claim: $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}_1 \cup agree_{\mathcal{S}}\mathcal{S}_2$. Proceeding to show it, recall from above that $agree^+_{\mathcal{S}}\mathcal{S}_\# \supseteq agree^+_{\mathcal{S}}\mathcal{S}_1 \cup agree^+_{\mathcal{S}}\mathcal{S}_2$ (this was item (1.) of the claim proven above). Notice that our definition of $\mathcal{S}'$ ensures that $agree^+_{\mathcal{S}}\mathcal{S}_\# \supseteq agree^+_{\mathcal{S}}\mathcal{S}'$: to obtain $\mathcal{S}'$, we have exclusively deleted pairs (w, v) from $\mathcal{S}_\#$ that are *not* in $\mathcal{S}$, in other words: for any pair $(w, v) \in \mathcal{S}_\#$ such that $(w, v) \in \mathcal{S}$, we have $(w, v) \in \mathcal{S}'$. It follows that $agree^+_{\mathcal{S}}\mathcal{S}' \supseteq agree^+_{\mathcal{S}}\mathcal{S}_1 \cup agree^+_{\mathcal{S}}\mathcal{S}_2$.

Next, suppose that $(w,v) \in agree^{-}_{\mathcal{S}}\mathcal{S}_i$, for some $i \in \{1,2\}$. This means that $(w,v) \notin \mathcal{S}$. Suppose first that $(w,v) \in agree^{-}_{\mathcal{S}}\mathcal{S}_{\#}$. Then $(w,v) \notin \mathcal{S}_{\#}$, and by definition of $\mathcal{S}'$, $(w,v) \notin \mathcal{S}'$, so $(w,v) \in agree^{-}_{\mathcal{S}}\mathcal{S}'$. Suppose, second, that $(w,v) \notin agree^{-}_{\mathcal{S}}\mathcal{S}_{\#}$. As we have seen above (item (2.) of the claim proven above), this implies that $w \approx_{\mathcal{S}_{\#}} v$. By definition of $\mathcal{S}'$, it follows that $(w,v) \notin \mathcal{S}'$, hence, again, $(w,v) \in agree^{-}_{\mathcal{S}}\mathcal{S}'$. It follows that $agree^{-}_{\mathcal{S}}\mathcal{S}' \supseteq agree^{-}_{\mathcal{S}}\mathcal{S}_1 \cup agree^{-}_{\mathcal{S}}\mathcal{S}_2$.

At this point, we are in a position to conclude that $agree_{\mathcal{S}}\mathcal{S}' \supseteq agree_{\mathcal{S}}\mathcal{S}_1 \cup agree_{\mathcal{S}}\mathcal{S}_2$, and the proof is complete. ⊣

This also completes the proof of Theorem 56. ⊣

This result allows us to prove canonicity results for a number of propositional attitudes, and dynamic attitudes realizing them, in a uniform manner: checking whether a given propositional attitude A satisfies our two properties, and finding a dynamic attitude τ which is optimal for A is sufficient to ensure that τ is actually canonical. Here is a sample of results:

COROLLARY 59.

1. *Infallible trust* $!$ *is canonical for irrevocable knowledge* K.
2. *Strong trust* $\Uparrow$ *is canonical for the disjunction of strong belief and opposite knowledge* $Sb \vee K^{\neg}$.
3. *Strong positive trust* $\Uparrow^{+}$ *is canonical for strong belief* Sb.
4. *Bare semi-trust* $!^{\sim}$ *is canonical for dual knowledge* $K^{\sim}$.
5. *Neutrality id is canonical for triviality* $\top$.
6. *Isolation* $\varnothing$ *is canonical for absurdity* $\bot$.

PROOF. For each of the six claims, it is easy to check that the propositional attitude A in question is preserved under refinements and admits joint embeddings. By Theorem 56, it follows that A is canonical. But if A is canonical, and τ is optimal for A, then τ is canonical for A. So we merely need to verify, for each item, that the τ in question is optimal for the A in question. For item (4.), this is easy to check; for item (1.), (2.), (5.) and (6.), cf. Proposition 37. For item (3.), the claim follows from Proposition 37 and the fact that, by Corollary 48, optimality is preserved under strictures. ⊣

Noticing that optimality is invariant under opposites, one obtains analogues results for the opposites of the attitudes mentioned in Corollary 59.

3.7. The Case of Simple Belief

We now turn to the second of the two questions raised towards the end of §3.4: do the failures of canonicity our formal theory gives rise to square with our intuitions about minimal change? Recall the observation that simple belief B is not canonical, which yields, as a consequence, that strict minimal trust $\uparrow^+$ is not canonical for B. This seems to be at odds with the special status Boutilier's minimal revision (i.e., essentially, our $\uparrow^+$) enjoys in the literature on belief revision. The operator $\uparrow^+$ realizes simple belief by making the best P-worlds the best worlds overall (if there are any, and otherwise deleting the whole order). One easily gets the sense that this is *the only reasonable thing to do* if, indeed, simple belief B is the target of revision. But our theory says otherwise: it allows many ways of realizing simple belief, all of them equally optimal, and none of them canonical. Our question was, essentially: is there anything we can do to improve this situation.

There is, of course, a question what is being evaluated, and what constitutes the yardstick of evaluation here. One reaction is to let the chips fall where they may: using our notion of canonicity as the yardstick, we could simply acknowledge that $\uparrow^+$ is not as special as we might have thought. But in this section, I want to take the other direction: using $\uparrow^+$, a prime example of what intuitively constitues a "natural" revision policy, as the yardstick, the question is: how to capture the sense in which this operator is unique?

This section explores three strategies to answer this question, which we will consider in sequence. Here is a preview: the *first strategy* (§3.7.1) is based on the intuition that receiving the information that P does not give us any reason to re-evaluate the plausibility hierarchy within the zone given by $P \cap \mathrm{S}$ within a plausibility order $\mathcal{S}$. If this is so, then the principle of informational economy suggests that we should, simply, keep the plausibility hierarchy within this zone the same. This idea can be implemented by means of restricting the class of dynamic attitudes we consider. The *second strategy* (§3.7.2) is, essentially, a variation on the first one: instead of *prohibiting* changes to the plausibility hierarchy among the P-worlds, we merely *discourage* them by imposing a penalty on such changes: other things being equal, an order $\mathcal{S}'$ that keeps the relative plausibility of P-worlds the same will count as more similar to the input order $\mathcal{S}$ than an order $\mathcal{S}''$ which does not keep it the same; that is: rather than forbidding changes among P-worlds, we merely flag them as "drastic". As we will see, both strategies solve our problem in the sense that they allow us to prove a uniqueness result for strict minimal trust $\uparrow^+$. Both strategies, however, also share the disadvantage that they seem to solve our problem by preempting it. For this reason, I tend to think of

the *third strategy* (3.7.3) explored below as the most insightful: it consists in refining our similarity measure with a positional component that essentially discourages moving worlds across larger distances than necessary.

3.7.1. Conservation. A dynamic attitude τ is *conserving* iff for any order $\mathcal{S}$, proposition P and worlds $w, v \in \mathrm{S}^{\tau P}$:

$$\text{if } w \in P \text{ iff } v \in P, \quad \text{then } w \leq_{\mathcal{S}} v \text{ iff } w \leq_{\mathcal{S}^{\tau P}} v.$$

Darwiche and Pearl, in their influential treatment of iterated revision, argued that conservation should be granted the status of a general postulate constraining belief revision, on a par with the original AGM postulates.[11]

But the property makes sense not only for operations on plausibility orders that induce (simple) belief, but as a general constraint on dynamic attitudes: one may justifiably wonder how obtaining information about P may give an agent any reason to reassess the relative plausibility of two P-worlds, or of two non-P-worlds. And if one draws the reasonable conclusion that the agent does not have any reason to do this, one will have arrived at the conclusion that dynamic attitudes should be conserving. The property is thus very natural, and, in fact, I am not aware of discussions of examples of *non-conserving* operations on plausibility orders.

Restricting attention to conserving attitudes is one way to solve our non-canonicity problem, as we show in Proposition 61 below.

lemma 60. *Let $\mathcal{S}$ be a plausibility order, suppose that P is substantial in $\mathcal{S}$, and let $\mathcal{S}' \in \mathrm{opt}_{\mathcal{S}} BP$. The following hold:*

1. $\mathrm{S}' = \mathrm{S}$.

2. *For all $w, v \in \mathrm{S}$: if $(w, v) \notin \mathrm{best}\, \mathcal{S}'$, then $(w, v) \in \mathcal{S}'$ iff $(w, v) \in \mathcal{S}$.*

proof. Let $\mathcal{S}$ be a plausibility order, suppose that P is substantial in $\mathcal{S}$, and let $\mathcal{S}' \in \mathrm{opt}_{\mathcal{S}} BP$. We consider the two items in turn.

1. Notice that $P \cap \mathrm{S} \neq \varnothing$, since P is, by assumption, substantial in $\mathcal{S}$. Now we know that $\mathrm{S}' \subseteq \mathrm{S}$. But if $\mathrm{S}' \subset \mathrm{S}$, then $\mathcal{S}^{\uparrow P} \prec_{\mathcal{S}} \mathcal{S}'$ (since $\mathrm{S}^{\uparrow P} = \mathrm{S}$), contradicting the assumption that $\mathcal{S}' \in \mathrm{opt}_{\mathcal{S}} BP$. So $\mathrm{S}' = \mathrm{S}$.

2. We define the order $\mathcal{S}''$ as follows:

   — $X_1 := \{(y, z) \in \mathcal{S}' \mid y \in \mathrm{best}\, \mathcal{S}'\}$,
   — $X_2 := \{(y, z) \in \mathcal{S} \mid y, z \notin \mathrm{best}\, \mathcal{S}'\}$,

[11] Cf. Darwiche and Pearl (1996).

— $\mathcal{S}'' := (S, X_1 \cup X_2)$.

Notice that, by definition of $\mathcal{S}''$, for all $w, v \in S$: if $(w,v) \notin \text{best}\,\mathcal{S}''$, then $(w,v) \in \mathcal{S}''$ iff $(w,v) \in \mathcal{S}$. In view of this observation, it is sufficient to show that $\mathcal{S}' = \mathcal{S}''$, as this implies our original claim. To prove that $\mathcal{S}' = \mathcal{S}''$, suppose that $(w,v) \in \mathit{agree}_{\mathcal{S}}\mathcal{S}'$. Suppose first that $w, v \notin \text{best}\,\mathcal{S}'$. Then $(w,v) \in \mathit{agree}_{\mathcal{S}}\mathcal{S}''$ by definition of $\mathcal{S}''$. Suppose, second, that it is not the case that $w, v \notin \text{best}\,\mathcal{S}'$. Then $(w,v) \in \mathcal{S}''$ iff $(w,v) \in \mathcal{S}'$ by definition of $\mathcal{S}'$, so also in this case, $(w,v) \in \mathit{agree}_{\mathcal{S}}\mathcal{S}''$. We conclude that $\mathcal{S}'' \preceq_{\mathcal{S}} \mathcal{S}'$. However, since $\mathcal{S}' \in \text{opt}_{\mathcal{S}}\, BP$, it is not the case that $\mathcal{S}'' \prec_{\mathcal{S}} \mathcal{S}'$. So $\mathcal{S}' = \mathcal{S}''$. As pointed out above, this implies our original claim, and we are done. ⊣

PROPOSITION 61. *Suppose that τ is conserving. If τ is optimal for belief, then $\tau = \uparrow^+$.*

PROOF. Suppose that τ is conserving and optimal for belief. Let $\mathcal{S}$ be a plausibility order, and let $P \subseteq W$. We discuss two cases. Suppose first that P is insubstantial in $\mathcal{S}$. Then $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^+ P}$, since $\overline{\tau} = \overline{\uparrow^+}$, so our claim holds. Suppose, second, that P is substantial in $\mathcal{S}$. Since τ is optimal, it follows by Proposition 47 that $\mathcal{S}^{\tau P} \in \text{opt}_{\mathcal{S}}\, BP$. We now make three observations:

1. By be the first item of the previous lemma, $S^{\tau P} = S$.
2. By the second item of the previous lemma, for all $w, v \in S$: if $w, v \notin \text{best}\,\mathcal{S}^{\tau P}$, then $(w,v) \in \mathcal{S}^{\tau P}$ iff $(w,v) \in \mathcal{S}$.
3. Since $\overline{\tau} = B$, it follows that $\text{best}\,\mathcal{S}^{\tau P} \subseteq P$. And since τ is conserving, it follows that for all $w, v \in S \cap P$: $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}} v$, hence $\text{best}\,\mathcal{S}^{\tau P} = \text{best}_{\mathcal{S}}\, P$.

Combining these observations, we conclude that $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^+ P}$. So, again, our claim holds. This shows that $\tau = \uparrow^+$ and completes the proof. ⊣

So assuming conservation as a background condition, we obtain the desired uniqueness result. In fact, our proof only uses a property that is weaker than conservation, namely the following one:

$$\forall w, v \in S^{\tau P} : \text{ if } w, v \in P, \text{ then } w \leq_{\mathcal{S}} v \text{ iff } w \leq_{\mathcal{S}^{\tau P}} v.$$

The solution of obtaining a uniqueness result by appeal to conservation does, however, look a little ad hoc. Assuming conservation, we are assuming that certain aspects of given structures are to be kept fixed. But the question *which* aspects of a structure are to be kept fixed is just what we are investigating. From this perspective, the solution seems to circumvent the problem, rather than addressing it at its core.

Keeping this in mind, we turn to the second of the three strategies suggested above.

3.7.2. Weighted Similarity. For any proposition P and plausibility orders $\mathcal{S}$ and $\mathcal{S}'$, let $\text{agree}_P(\mathcal{S},\mathcal{S}') := \text{agree}(\mathcal{S},\mathcal{S}') \cap (P \times P)$. Given a plausibility order $\mathcal{S}$, and a proposition P, we define the order $(O_\mathcal{S}, <^P_\mathcal{S})$ as follows:

$\mathcal{S}' <^P_\mathcal{S} \mathcal{S}''$ iff

— $\mathrm{S}' \supset \mathrm{S}''$, *or*
— $\mathrm{S}' = \mathrm{S}''$ and $\text{agree}^P_\mathcal{S}\,\mathcal{S}' \supset \text{agree}^P_\mathcal{S}\,\mathcal{S}''$, *or*
— $\mathrm{S}' = \mathrm{S}''$ and $\text{agree}^P_\mathcal{S}\,\mathcal{S}' = \text{agree}^P_\mathcal{S}\,\mathcal{S}''$ and $\text{agree}_\mathcal{S}\,\mathcal{S}' \supset \text{agree}_\mathcal{S}\,\mathcal{S}''$.

In comparing the similarity of two orders $\mathcal{S}'$ and $\mathcal{S}''$ to a given order $\mathcal{S}$, the above definition ensures that, other things being equal, a penalty is imposed for changing the relative plausibility of P-worlds.

Call a dynamic attitude τ *weighted optimal* if there exists no attitude σ and plausibility order $\mathcal{S}$ such that $\overline{\sigma} = \overline{\tau}$ and $\mathcal{S}^{\sigma P} <^P_\mathcal{S} \mathcal{S}^{\tau P}$; and call τ *weighted optimal for* A if τ is weighted optimal and $\overline{\tau} = A$.

proposition 62. *If τ is weighted optimal for simple belief B, then $\tau = \uparrow^+$.*

proof. Suppose that τ is weighted optimal for simple belief. Let $\mathcal{S}$ be a plausibility order, and P a proposition. If P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^+ P}$. Suppose now that P is substantial in $\mathcal{S}$. We have three observations to make:

1. $\mathrm{S}^{\tau P} = \mathrm{S}$ (for otherwise, $\mathcal{S}^{\uparrow^+ P} <^P_\mathcal{S} \mathcal{S}^{\tau P}$, contradiction).

2. $\text{best}\,\mathcal{S}^{\tau P} = \text{best}_\mathcal{S}\,P$. This follows from the fact that $\mathcal{S}^{\tau P} \models BP$, together with the fact that for any $w, v \in P \cap \mathrm{S}$: $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}^{\uparrow^+ P}} v$ (for otherwise, $\mathcal{S}^{\uparrow^+ P} <^P_\mathcal{S} \mathcal{S}^{\tau P}$, contradiction).

3. For all $w, v \in \mathrm{S}$ such that $w, v \notin \text{best}\,\mathcal{S}^{\tau P}$: $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_\mathcal{S} v$ (for otherwise, $\mathcal{S}^{\uparrow^+ P} <^P_\mathcal{S} \mathcal{S}^{\tau P}$, contradiction).

Taken together, these observations imply that $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^+ P}$. So regardless of whether P is substantial or insubstantial in $\mathcal{S}$, we have shown that $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^+ P}$. This proves our initial claim. ⊣

So again, we obtain a uniqueness result, as desired. The main criticism that may be adduced against this solution to our problem is that in working with weighted similarity, we are introducing a new parameter on which similarity comparisons depend. Given two plausibility orders $\mathcal{S}', \mathcal{S}'' \in O_\mathcal{S}$, the question which of the two is "more similar" to $\mathcal{S}$ has no immediate answer anymore, as we need to know which proposition P is used as a criterion of comparison. This seems undesirable. And indeed, as we show in the next paragraph, we do not really need the additional parameter to obtain a uniqueness result for simple belief. Adding a *positional* component to the notion of similarity works just as well.

3.7.3. Positional Similarity. The way I will think about "positions" of worlds in this section is that a world *w maintains its position* in a ("transformed") order $\mathcal{S}'$ compared to an ("initial") order $\mathcal{S}$ if, in going from $\mathcal{S}$ to $\mathcal{S}'$, no world gets advanced to the same plausibility level as w, and no world gets promoted across the plausibility level of w. Taken together, this indicates that the position of w is "at least as good" in $\mathcal{S}'$ as it used to be in $\mathcal{S}$.

Formally, let $\mathcal{S}, \mathcal{S}'$ be plausibility orders such that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, and let $w \in \mathrm{S}'$. We define the proposition $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$ by stipulating that $w \in \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$ iff for all $v \in \mathrm{S}'$:

— If $v \approx_{\mathcal{S}'} w$, then $v \approx_{\mathcal{S}} w$
— If $v <_{\mathcal{S}'} v$, then $v <_{\mathcal{S}} w$.

Note that membership of w in $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$ does not exclude that w itself gets advanced relative to other worlds. However, this change will be recorded *by those other worlds* (which will fail to maintain their position), rather than by w itself.

On the basis of our formal conception of what it means for a world to maintain position, we now define the following measure of similarity, which we denote $<^{\circ}_{\mathcal{S}}$ (the $\circ$ symbol is supposed to remind the reader of the word "position"), given an order $\mathcal{S}$, putting $\mathcal{S}' <^{\circ}_{\mathcal{S}} \mathcal{S}''$ iff:

— $\mathrm{S}'' \subset \mathrm{S}'$, *or*
— $\mathrm{S}'' = \mathrm{S}'$ and $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'' \subset \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$ *or*
— $\mathrm{S}'' = \mathrm{S}'$ and $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'' = \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$ and $\mathrm{agree}_{\mathcal{S}}\, \mathcal{S}'' \subset \mathrm{agree}_{\mathcal{S}}\, \mathcal{S}'$.

To get a feeling how this notion of similarity relates to our problem, we consider two examples. Adopting the notation of Figure 18, we have that $\mathcal{S}' <^{\circ}_{\mathcal{S}} \mathcal{S}''$, since $\mathrm{S}' = \mathrm{S}''$ and $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'' = \{y\} \subset \{x, y\} = \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$. Adopting now the notation of Figure 19, we notice again that $\mathcal{S}' <^{\circ}_{\mathcal{S}} \mathcal{S}''$, since $\mathrm{S}' = \mathrm{S}''$ and $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'' = \{y\} \subset \{x, y\} = \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$, for the same reason: $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'' = \{y\} \subset \{x, y\} = \mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}'$.

The reader may notice that, under the assumption that $x, y \in P$ and $w \notin P$, $\mathcal{S}'$ actually equals $\mathcal{S}^{\uparrow^+ P}$, both in Figure 18 and in Figure 19.

I believe that it is fairly easy to intuitively grasp that $\uparrow^+$ has to be the *only* dynamic attitude that is positionally optimal for belief. But going through all the details requires some work. We do all the preparations in Lemma 63–65 below, while Proposition 66 records the desired result.

lemma 63. *Let $\mathcal{S}, \mathcal{S}'$ be a plausibility order, let P be substantial in $\mathcal{S}$ and suppose that $\mathcal{S}' \models BP$. Then:*

1. $\mathrm{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P} = \mathrm{best}_{\mathcal{S}}\, P \cup \{v \in \mathrm{S} \mid \exists w \in \mathrm{best}_{\mathcal{S}}\, P : w <_{\mathcal{S}} v\}$.

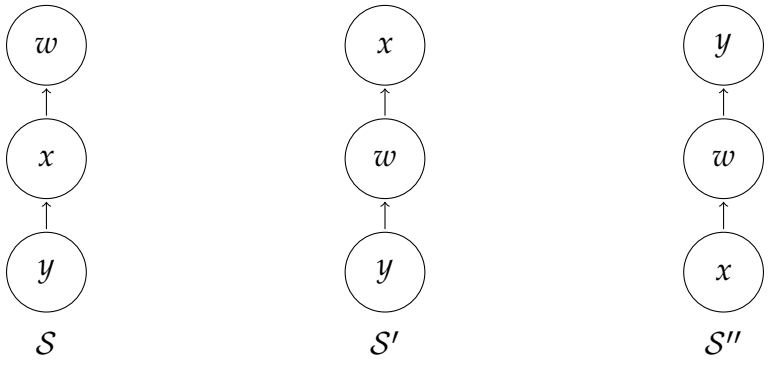


FIGURE 18. Going from $\mathcal{S}$ to $\mathcal{S}'$, the world x is promoted over all other worlds, while going from $\mathcal{S}$ to $\mathcal{S}''$, the world y is promoted over all other worlds.

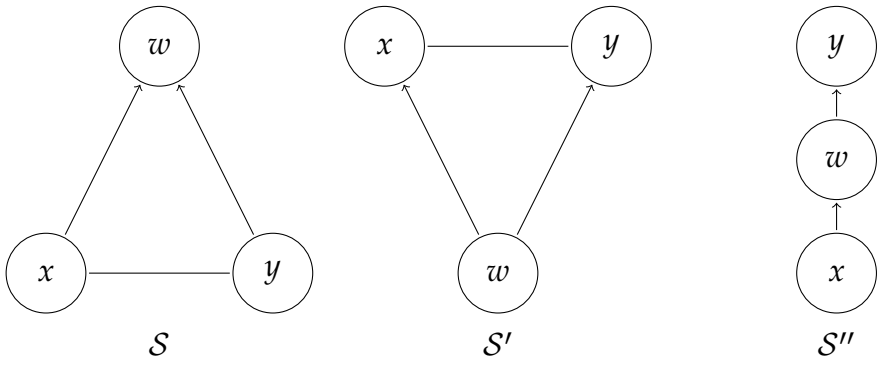


FIGURE 19. Going from $\mathcal{S}$ to $\mathcal{S}'$, both x and y are promoted over w, while going from $\mathcal{S}$ to $\mathcal{S}''$, only y is promoted (over both w and x).

2. $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' \subseteq \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$.

3. *If* $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$, *then* $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' = \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$.

PROOF. The *first item* is immediate by definition of $\uparrow$. As for the *second item*, towards a contradiction, suppose that $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' \not\subseteq \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. Then there exists $x \in \text{maintain}_{\mathcal{S}}\, \mathcal{S}'$ such that $x \notin \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. By the first item, $x <_{\mathcal{S}} y$ for any $y \in P \cap \text{S}$. So $x \notin P$. Since $x \in \text{maintain}_{\mathcal{S}}\, \mathcal{S}'$, it follows that for any $y \in P \cap \text{S}$: $x <_{\mathcal{S}'} y$ (for otherwise, $y \leq_{\mathcal{S}} x$, contradiction). It follows that $\text{best}\, \mathcal{S}' \not\subseteq P$, so $\mathcal{S}' \not\models BP$. This contradicts our initial assumption, and the claim follows, which finishes the second item. For the *third item*, suppose that $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$. Notice that if $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' \subseteq \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$, it follows that $\mathcal{S}^{\uparrow P} <^{\circ}_{\mathcal{S}} \mathcal{S}'$, which contradicts the assumption that $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$. Hence, by the

second item, the claim follows: $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' = \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. ⊣

LEMMA 64. *Let P be substantial in $\mathcal{S}$, suppose that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, and let $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$.*

1. $\text{S}' = \text{S}$.

2. $\text{best}\, \mathcal{S}' = \text{best}_{\mathcal{S}}\, P$.

PROOF. Let P be substantial in $\mathcal{S}$, and let $\mathcal{S}' \in \text{opt}^{2}_{\mathcal{S}}\, BP$. We consider the three items in turn.

1. We know that $\text{S}' \subseteq \text{S}$. Suppose that $\text{S}' \subset \text{S}$. Then $\mathcal{S}^{\uparrow P} <^{2}_{\mathcal{S}} \mathcal{S}'$, so $\mathcal{S}' \notin \text{opt}^{\circ}_{\mathcal{S}}\, BP$, contradiction. Hence $\text{S}' = \text{S}$.

2. We consider the two halves of the claim in turn and show them by reductio. For one half, suppose that $\text{best}_{\mathcal{S}}\, P \nsubseteq \text{best}\, \mathcal{S}'$. Choose a world $x \in \text{S}$ such that $x \in \text{best}_{\mathcal{S}}\, P$, $x \notin \text{best}\, \mathcal{S}'$. Since $\mathcal{S}' \vDash BP$, we have that $\text{best}\, \mathcal{S}' \subseteq P$, so there exists $w \in P \cap \text{S}$: $y <_{\mathcal{S}'} x$. Since $x \in \text{best}_{\mathcal{S}}\, P$, it follows that $x \leq_{\mathcal{S}} y$. So $x \notin \text{maintain}_{\mathcal{S}}\, \mathcal{S}'$. By the previous lemma (first item), $x \in \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. So $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' \neq \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. However, again by the previous lemma (third item), $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' = \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. This is a contradiction. We conclude that $\text{best}\, \mathcal{S}' \subseteq \text{best}_{\mathcal{S}}\, P$.

   For the other half, suppose that $\text{best}\, \mathcal{S}' \nsubseteq \text{best}_{\mathcal{S}}\, P$. Choose a world $x \in \text{S}$ such that $x \in \text{best}\, \mathcal{S}'$, $x \notin \text{best}_{\mathcal{S}}\, P$. Since $\mathcal{S}' \vDash BP$, it follows that $x \in \text{S} \cap P$. Since $x \notin \text{best}_{\mathcal{S}}\, P$, there exists a world $y \in \text{best}_{\mathcal{S}}\, P$ such that $y <_{\mathcal{S}} x$. Since $x \in \text{best}\, \mathcal{S}'$, we have $x \leq_{\mathcal{S}'} y$. So $y \notin \text{maintain}_{\mathcal{S}}\, \mathcal{S}'$. Now we argue as in the first half of the proof for this item, and arrive at the desired result: $\text{best}_{\mathcal{S}}\, P \subseteq \text{best}\, \mathcal{S}'$. ⊣

LEMMA 65. *Let P be substantial in $\mathcal{S}$, suppose that $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, and let $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$. Then $\mathcal{S}' = \mathcal{S}^{\uparrow P}$.*

PROOF. We first observe that, since $\text{S}' = \text{S}^{\uparrow P} = \text{S}$ (first item of Lemma 64) and $\text{maintain}_{\mathcal{S}}\, \mathcal{S}' = \text{maintain}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$ (third item of Lemma 63), the following holds by definition of $<^{\circ}_{\mathcal{S}}$:

$$\text{If } \text{agree}_{\mathcal{S}}\, \mathcal{S}' \subset \text{agree}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P} \text{ then } \mathcal{S}^{\uparrow P} <^{\circ}_{\mathcal{S}} \mathcal{S}'.$$

It is thus sufficient to show that $\text{agree}_{\mathcal{S}}\, \mathcal{S}' \subseteq \text{agree}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$, from which we may conclude that, actually, $\text{agree}_{\mathcal{S}}\, \mathcal{S}' = \text{agree}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$ (for otherwise, we obtain a contradiction to the assumption that $\mathcal{S}' \in \text{opt}^{\circ}_{\mathcal{S}}\, BP$). But from $\text{agree}_{\mathcal{S}}\, \mathcal{S}' = \text{agree}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$ it follows that $\mathcal{S}' = \mathcal{S}^{\uparrow P}$.

To show the claim, we establish that for any $(w, v) \in \text{S} \times \text{S}$: if $(w, v) \in \text{agree}_{\mathcal{S}}\, \mathcal{S}'$, then $(w, v) \in \text{agree}_{\mathcal{S}}\, \mathcal{S}^{\uparrow P}$. Letting $(w, v) \in \text{S} \times \text{S}$, we discuss two cases.

Case 1: If at least one of w and v is an element of $\mathrm{best}\,\mathcal{S}'$, our claim holds, for in that case, $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}'$ iff $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow P}$ using the definition of $\uparrow$ and the fact that $\mathrm{best}\,\mathcal{S}' = \mathrm{best}\,\mathcal{S}^{\uparrow P}$ (Lemma 64). *Case 2:* Suppose that $w, v \notin \mathrm{best}\,\mathcal{S}'$. By definition of $\uparrow$, $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow P}$. Thus if $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}'$, then $(w,v) \in \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow P}$.

This is the desired result, so we may, indeed, conclude that $\mathrm{agree}_{\mathcal{S}}\,\mathcal{S}' = \mathrm{agree}_{\mathcal{S}}\,\mathcal{S}^{\uparrow P}$, and thus $\mathcal{S}' = \mathcal{S}^{\uparrow P}$. ⊣

PROPOSITION 66. *Let τ be a dynamic attitude and suppose that τ is positionally optimal for belief. Then $\tau = \uparrow^{+}$.*

PROOF. Let $\mathcal{S}$ be a plausibility order, and $P \subseteq W$. If P is insubstantial in $\mathcal{S}$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow^{+} P}$ since $\overline{\tau} = \overline{\uparrow^{+}}$. If P is substantial in $\mathcal{S}$, then, observing that $\mathcal{S}^{\tau P} \in \mathrm{opt}^{\circ}_{\mathcal{S}}\, BP$ by our initial assumption, we apply Lemma 65 to conclude that $\mathcal{S}^{\tau P} = \mathcal{S}^{\uparrow P} = \mathcal{S}^{\uparrow^{+} P}$. It follows that $\tau = \uparrow^{+}$. ⊣

Notice that this result provides, after all, a motivation for the claim that dynamic attitudes realizing simple belief should be conserving in order to adhere to the principle of minimal change: $\uparrow^{+}$ emerges as the *only* dynamic attitude that is positionally optimal for belief. And $\uparrow^{+}$ is conserving. In this sense, the previous result may be seen as a (conceptual) improvement on Proposition 61, according to which the only conserving dynamic attitude that is optimal for belief is $\uparrow^{+}$. There, we were assuming conservation as a background property. Here, it falls out as a consequence of more general considerations.

3.7.4. (WEAK) SEMI-TRUST. Finally, we notice that the notion of positional similarity also allows us to prove uniqueness results for the dynamic attitudes $\Uparrow^{\sim}$ (weak semi-trust) and $\uparrow^{\sim}$ (semi-trust).

PROPOSITION 67. *Let τ be a dynamic attitude.*

— *Suppose that τ is positionally optimal for dual belief $B^{\sim}$. Then $\tau = \uparrow^{\sim +}$.*
— *Suppose that τ is positionally optimal for dual strong belief $Sb^{\sim}$. Then $\tau = \Uparrow^{\sim +}$.*

As the proof is similar to the one given for $\uparrow$, we do not provide details here.

Chapter 4. Robustness

In the preceding chapters, we have studied dynamic attitudes in terms of the propositional attitudes they realize as their targets. In this chapter, we change the perspective, considering the *robustness* (or *stability*) of propositional attitudes under various kinds of transformations. This issue is usually discussed under the heading of "preservation": one investigates the properties of given structures that are preserved (or remain stable) under particular operations.

Preservation is an important topic in model theory—typical results in this area characterize model-theoretic transformations in terms of classes of sentences (i.e., properties, given a semantics for these sentences) that they preserve. The Łoś-Tarski Theorem is perhaps the most famous result in this category, showing that the first-order sentences preserved under taking substructures are just the universal ones.[1] Preservation has also been an important topic in dynamic semantics. Here, the question which sentences of a given language are *persistent* arises frequently: which sentences φ have the property that support for φ is, in general, stable under updates with arbitrary further sentences?[2] In epistemology, the stability of knowledge has been an important concern, with Lehrer (1990)'s defeasibility theory being a prominent example (roughly, according to Lehrer's theory: "one knows something if one's commitment to it is stable under influx of arbitrary true information").[3] In belief revision, preservation questions also play a role, as we will see below, even if perhaps in an implicit manner.

Our approach in this chapter has a dynamic twist: we aim to characterize particular classes of propositional attitudes—for example, the ones that are preserved under substructures—in terms of classes of dynamic attitudes realizing the former. The main results of this chapter—Theorem 74, 75 and

[1] Cf., e.g., Hodges (1997).

[2] Persistence plays, for example, an important role in the work of Groenendijk, Stokhof, and Veltman (1996), Veltman (2005), Gillies (2010), Willer (2012).

[3] Cf. also Rott (2004), Stalnaker (2006), Baltag and Smets (2008).

76—are of this kind.

§4.1 shows how preservation questions naturally arise in the existing literature from belief revision theory; §4.2 provides characterizations of the propositional attitudes that are *persistent* (stable under any transformation given by an upgrade) and of the propositional attitudes that are *preserved under substructures* (stable under upgrade that deletes worlds while keeping the order on the remaining worlds the same); in the process, we identify an important subclass of dynamic attitudes whose fixed points are preserved under substructures: the *distributive* dynamic attitudes; §4.3 and §4.4 investigate this class further. In particular, our analysis yields a novel characterization of strong trust $\Uparrow$ (lexicographic upgrade) and infallible trust $!$ (update).

4.1. A Puzzle in Iterated Revision

In a famous paper, Darwiche and Pearl (1996) discuss the following example:

> *We encounter a strange new animal and it appears to be a bird, so we believe the animal is a bird. As it comes close to our hiding place, we see clearly that the animal is red, so we believe that it is a red bird. To remove further doubts about the animal birdness, we call in a bird expert who takes it for examination and concludes that it is not really a bird but some sort of mammal. The question now is whether we should still believe that the animal is red.*

Darwiche and Pearl claim that the answer to the above question should be an unqualified "Yes." They write: "once the animal is seen red, it should be presumed red no matter what ornithological classification it obtains." Other authors, for example, Booth and Meyer (2006), have gone on the record as sharing the intuition; it is also shared by the present author. Here, I want to explore the consequences of accepting the intuition from the perspective of the topic of this chapter.

The reason the example is interesting in the present context is that Darwiche and Pearl are, essentially, discussing a *preservation question*: Darwiche and Pearl claim that our agent, upon observing that an animal, believed to be a bird, is red, should acquire a belief in the redness of that animal that is *stable* (or *preserved*) under revision with the observation that the animal is not, in fact, a bird.

4.1.1. The Problem. To get clearer about what is going on, let us first see in how far our scenario presents a *problem*. We start with a more rigorous

model of the situation. The first assumption we are going to make is that all relevant information is encoded in the story as presented in the scenario. As in other places, this is also crucial here. In particular, it means that the properties of "being red" and "being a bird" are to be construed as *unrelated* from the perspective of our agent. Consider a variant: we observe an animal and conclude that it is not a tiger. As it comes closer, we become convinced that it is not dangerous, so it's a non-dangerous non-tiger. But then, some expert convinces us that the animal is, in fact, a tiger. Would we then believe that the animal (a tiger!) is not dangerous? Perhaps not, since, after all: tigers can be dangerous. So the intuition described by Darwiche and Pearl rests on the background assumption that an animal's not being a bird does not, from the perspective of our agent, count as evidence against that animal's being red.

So to make sure that no other background assumptions slip in, we start from the assumption that the agent has no prior evidence as to the relative plausibility of worlds in which the animal is red, and worlds in which the animal is a bird.

This leads to the plausibility order depicted in Figure 20. It consists of four equiplausible possible worlds: the animal could be a red bird, a red non-bird, a non-red bird, or a non-red non-bird.

Now what is the problem? The answer is in Figure 21: if we upgrade the initial model with a series of minimal upgrades—thus assuming that the agent places minimal trust $\uparrow$ in the sources from which she receives information—, the agent will, after receiving the three pieces of information in turn, not believe that the animal is red. Whoever accepts the intuition presented by Darwiche and Pearl will thus have problems in accepting minimal trust $\uparrow$ as a good model for an agent's belief revision policy. And since in our example, only *substantial* propositions are involved (propositions that are neither known to be false nor known to be true upfront), this criticism extends with the same force to *strict minimal trust* $\uparrow^+$.

Note that this is explicitly a criticism of strict minimal trust, but implicitly a criticism of the *fixed point* of strict minimal trust, which is simple belief. So what Darwiche and Pearl are really saying here is that simple belief is not the appropriate propositional target attitude to model an agent's belief revision processes. And the reason for this is that simple belief fails to be *robust* enough, i.e., it fails to satisfy a preservation property that it intuitively should satisfy.

4.1.2. Two Possible Solutions. Let us first observe that there are *solutions* to the problem: instead of minimal trust (or strict minimal trust), one may

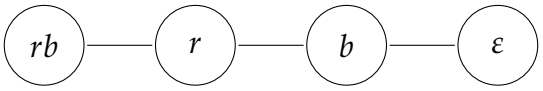


FIGURE 20. The plausibility order $\mathcal{S}$ representing the initial situation in the Darwiche-Pearl scenario. Here, rb is the world in which the animal is a red bird, r is the world in which the animal is a red non-bird, b is the world in which the animal is a non-red bird, and ε is the world in which the animal is a non-red non-bird.

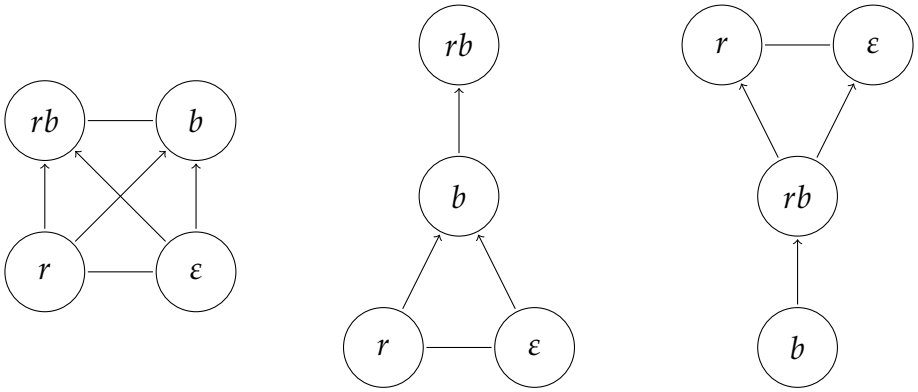


FIGURE 21. A sequence of minimal upgrades applied to the plausibility order $\mathcal{S}$ depicted in Figure 20 above. From left to right: $\mathcal{S}^{\uparrow B}$, $(\mathcal{S}^{\uparrow B})^{\uparrow R}$, $((\mathcal{S}^{\uparrow B})^{\uparrow R})^{\uparrow \neg B}$.

choose to work with other dynamic attitudes that escape the criticism by Darwiche and Pearl. We depict two of them in Figure 22 and Figure 23, respectively. As the reader may want to check, if we use moderate trust $\Uparrow$ (as in Figure 22) or strong trust $\Uparrow$ (as in Figure 23), it is indeed the case that after the sequence of upgrades in our scenario, applied to the initial plausibility order $\mathcal{S}$ (depicted in Figure 20), the agent believes that the animal in question is red.[4]

The propositional attitude towards the proposition R that is created by $\Uparrow$ and $\Uparrow$, respectively, is thus *stable* (or *robust*) enough to be preserved under the subsequent upgrade with the proposition $\neg B$.

As an aside, notice one aspect in which the two solutions are interestingly different. In the order $((\mathcal{S}^{\Uparrow B})^{\Uparrow R})^{\Uparrow \neg B}$ depicted on the right-hand side of Figure 22, the agent has a *refined belief* that the animal is red; however, in the order

[4]The first observation is due to Darwiche and Pearl (1996), and the second is due to Booth and Meyer (2006).

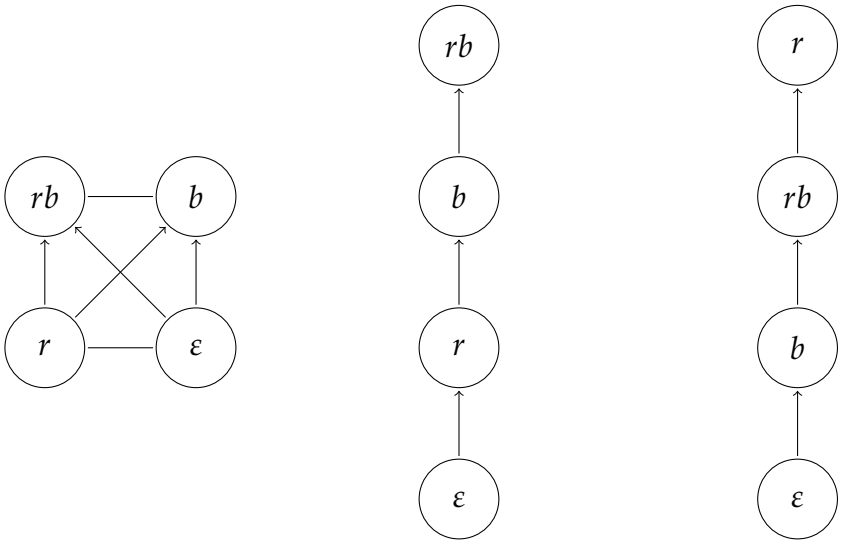


FIGURE 22. A sequence of moderate upgrades applied to the plausibility order $\mathcal{S}$ depicted in Figure 20 above. From left to right: $\mathcal{S}^{\Uparrow B}$, $(\mathcal{S}^{\Uparrow B})^{\Uparrow R}$, $((\mathcal{S}^{\Uparrow B})^{\Uparrow R})^{\Uparrow \neg B}$.

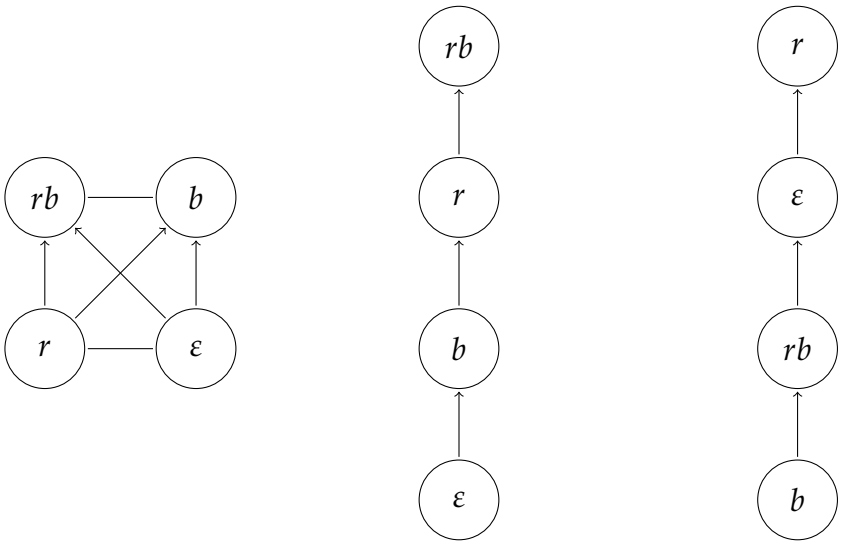


FIGURE 23. A sequence of lexicographic upgrades applied to the plausibility order $\mathcal{S}$ depicted in Figure 20 above. From left to right: $\mathcal{S}^{\Uparrow B}$, $(\mathcal{S}^{\Uparrow B})^{\Uparrow R}$, $((\mathcal{S}^{\Uparrow B})^{\Uparrow R})^{\Uparrow \neg B}$.

$((\mathcal{S}^{\Uparrow B})^{\Uparrow R})^{\Uparrow \neg B}$, the agent does *not* have a *strong belief* that the animal is red. So refined beliefs are preserved under applying moderate trust in circumstances

where strong beliefs are not preserved under applying strong trust.

These are but two of potentially many solutions to the problem. As pointed out above, the puzzle discovered by Darwiche and Pearl has the distinctive flavour of a preservation question. It seems that the *robustness, or lack of robustness, of the fixed point of the dynamic attitude that we use* is the factor determining whether the problem diagnosed by Darwiche and Pearl arises. Our analysis so far merely shows that the fixed point of $\uparrow$ is not robust enough, while the fixed point of $\Uparrow$ and $\Uparrow\Uparrow$ is. Our goal is now to clarify what preservation property is exactly at stake here.

4.1.3. DP-Robustness. In analyzing the example from a more general perspective, we can simplify matters somewhat by generalizing its structure. Consider the following variant of the scenario:

> *We encounter a strange new animal. As it comes close to our hiding place, we see clearly that the animal is red, so we believe that it is red. To determine whether it is a bird, we call in a bird expert who takes it for examination and concludes that it is not really a bird but some sort of mammal. The question now is whether we should still believe that the animal is red.*

The intuition, I submit, is as clear as in the original scenario: we should continue to believe that the animal is red. The intuition is thus independent of the initial belief that the animal is a bird (which is later overruled by the bird expert). Any solution to the puzzle posed by Darwiche and Pearl needs to work just as well in the modified scenario.

What matters, however, is the following: in the modified scenario, as in the original one, there is no initial dependency among the two propositions in the sense that coming to *know* that the animal is not a bird would be sufficient to convince us that the animal is not red. More precisely, among the most plausible worlds in which the animal is not a bird there are worlds where the animal is red.

We are thus envisaging a plausibility order $\mathcal{S}$ such that

$$\mathcal{S} \vDash B^{\sim\neg B} R$$

("upon obtaining the hard information that the animal is a bird, it could still be red").[5] Our intuition now boils down to the fact that after accepting, first,

[5]Note that we do not exclude that, in $\mathcal{S}$, our agent actually believes the animal to be a bird: we are merely factoring out the assumption, by allowing the agent to be opinionated or unopinionated—our solution has to work regardless.

that the animal is red, and second, that the animal is *not* a bird, the agent should *still* believe the animal to be red. That is, we would like to have a dynamic attitude τ that (1) is positive, and (2) satisfies

$$((\mathcal{S})^{\tau Q})^{\tau P} \vDash BQ.$$

Call this requirement (†).

Our observation is now that we can, if we restrict attention—as Darwiche and Pearl do—to *conserving* dynamic attitudes, characterize these two requirements in terms of a property that clearly brings out the connection to our topic: preservation.

We call a positive dynamic attitude τ *DP-robust* if τ creates belief and for any plausibility order $\mathcal{S}$ and propositions P and Q, the following holds:

$$\text{If } \mathcal{S} \vDash B^{\sim P}Q, \text{ then } \mathcal{S}^{\tau Q} \vDash BQ \wedge B^{P}Q.$$

Equivalently, τ is DP-robust if τQ creates, whenever the above antecedent is satisfied, a belief in Q that is preserved (!) under coming to know that P from an infallible source, i.e.:

$$\text{If } \mathcal{S} \vDash B^{\sim P}Q, \text{ then } \mathcal{S}^{\tau P} \vDash BQ \text{ and } (\mathcal{S}^{\tau Q})^{!P} \vDash BQ.$$

The antecedent of the conditional matches the background assumption in the Darwiche-Pearl scenario: initially, coming to know that the animal is not a bird is not conclusive evidence against the animal's being red, that is, in the initial situation of the scenario given by the plausibility order $\mathcal{S}$ depicted in Figure 20, we have that $\mathcal{S} \vDash B^{\sim\neg B}R$. The above condition then requires that both $\mathcal{S}^{\tau P} \vDash BR$ and $(\mathcal{S}^{\tau R})^{!\neg B} \vDash BR$.

As the next observation shows, the above requirement (†) and the notion of DP-robustness coincide under the assumption that τ is conserving:

PROPOSITION 68. *Let τ be a positive conserving dynamic attitude. The following are equivalent:*

— *τ is DP-robust.*
— *For any plausibility order $\mathcal{S}$ such that $\mathcal{S} \vDash B^{\sim P}Q$: $((\mathcal{S})^{\tau Q})^{\tau P} \vDash BQ$.*

PROOF. From (1.) to (2.), we argue as follows: suppose that τ is DP-robust. Further, since $\mathcal{S} \vDash B^{\sim P}Q$, it follows from the assumption that $\mathcal{S}^{\tau Q} \vDash B^{P}Q$. Let $\mathcal{T} = (\mathcal{S}^{\tau Q})^{\tau P}$. Since τ is conserving, $\text{best}\,\mathcal{T} = \text{best}_{\mathcal{S}^{\tau Q}} P$. Since $\mathcal{S}^{\tau Q} \vDash B^{P}Q$, it follows that $\text{best}_{\mathcal{S}^{\tau Q}} P \subseteq Q$. Hence $\mathcal{T} \vDash BQ$.

From (2.) to (1.), we argue as follows. Let $\mathcal{S}$ be a plausibility order and suppose that $\mathcal{S} \vDash B^{\sim P}Q$. By the assumption, $\mathcal{T} = ((\mathcal{S})^{\tau Q})^{\tau P} \vDash BQ$. Hence $\text{best}\,\mathcal{T} \subseteq Q$ (by definition of belief). Since τ is conserving, $\text{best}_{\mathcal{S}^{\tau Q}} P = \text{best}\,\mathcal{T}$. So $\text{best}_{\mathcal{S}^{\tau Q}} P \subseteq Q$. Hence $\mathcal{S}^{\tau Q} \vDash B^{P}Q$. ⊣

The result clarifies the connection between the original Darwiche-Pearl scenario and preservation questions. A natural class of DP-robust dynamic attitudes is now found among the moderately positive attitudes (cf. §2.5, and in particular, Table 2).

We call a dynamic attitude τ *weakly soft* if $\mathrm{S} \cap P = \varnothing \Rightarrow \mathrm{S}^{\tau P} = \mathrm{S}$ for any proposition P and plausibility order $\mathcal{S}$. Weakly soft attitudes are those dynamic attitudes that do not provide hard information in case their propositional argument is consistent with what the agent already knows. Unless the agent happens to receive information that is absurd against the background of his knowledge, the information provided by a weakly soft attitude is thus defeasible. This property is implicit in the setting presented by Darwiche and Pearl, where *all* belief revision policies are assumed to be weakly soft (contrast this with our general notion of a dynamic attitude, which does allow an increase of hard information due to an upgrade).

PROPOSITION 69. *Let τ be a weakly soft dynamic attitude. If τ is moderately positive, then τ is DP-robust.*

PROOF. Suppose that τ is a weakly soft, moderately positive dynamic attitude. Assume that $\mathcal{S} \vDash B^{\sim P}Q$. We have to show that $\mathcal{S}^{\tau Q} \vDash B^{P}Q$, or, equivalently, that $\mathrm{best}_{\mathcal{S}^{\tau Q}} P \subseteq Q$. Towards a contradiction, suppose that there exists $w \in \mathrm{best}_{\mathcal{S}^{\tau Q}} P$ such that $w \notin Q$. By the assumption that $\mathcal{S} \vDash B^{\sim P}Q$, we know that there exists $v \in \mathrm{best}_{\mathcal{S}} P$ such that $v \in Q$ and $w \approx_{\mathcal{S}} v$. Since τ is weakly soft, $v \in \mathrm{S}^{\tau P}$. Since τ is moderately positive, $v <_{\mathcal{S}^{\tau P}} w$. Hence $w \notin \mathrm{best}_{\mathcal{S}^{\tau Q}} P$. This is a contradiction. So $\mathrm{best}_{\mathcal{S}^{\tau Q}} P \subseteq Q$. It follows that τ is DP-robust. ⊣

Hence:

COROLLARY 70. *Let τ be a weakly soft conserving dynamic attitude. Then for any plausibility order $\mathcal{S}$ such that $\mathcal{S} \vDash B^{\sim P}Q$: $((\mathcal{S})^{\tau Q})^{\tau P} \vDash BQ$.*

PROOF. Immediate from Proposition 68 and Proposition 69. ⊣

This result gives us some insight into why moderate trust $\Uparrow$ and strong trust $\Uparrow$ are solutions to the problem raised by Darwiche and Pearl (as pointed out in §4.1.2 above): they are solutions in virtue of the fact that they are both weakly soft and moderately positive on the one hand, and thus DP-robust, and conserving on the other hand.

The purpose of the present section was to show that preservation questions arise naturally in existing research in belief revision theory. As a preservation property, DP-robustness is already of a rather complex nature. In the next section, we continue our investigation of the topic of preservation of propositional attitudes, while focusing on a much simpler property that has traditionally been important in model theory: preservation under substructures.

4.2. *Preservation under Substructures*

The preservation property considered in the previous section had a rather complex structure. Motivated by our example, we were interested in the question whether a particular positive dynamic attitude τ has the property that the *belief* induced by an upgrade τQ applied to an initial order $\mathcal{S}$ is preserved under performing certain upgrades τP, namely upgrades with propositions P such that Q was not implausible given P.

We now move to a less fine-grained stance, asking, more simply, whether a particular propositional attitude is preserved under certain operations on a plausibility order in which it is satisfied. In addressing this question, we will maintain, however, our dynamic perspective on the matter: since propositional attitudes arise as fixed points of dynamic attitudes, we may strive for a dynamic characterization of the propositional attitudes that are preserved under particular operations. Once again, the link is provided by our notion of a *fixed point*.

We will be interested in two preservation properties in this section: *preservation under substructures* and *persistence*. As pointed out at the beginning of this chapter, these have traditionally played an important role, in dynamic semantics as well as in model theory.

— Let $\mathcal{S}$, $\mathcal{S}'$ be plausibility orders. $\mathcal{S}'$ is a *substructure* of $\mathcal{S}$ if there exists a proposition Q such that $\mathcal{S}|_Q = \mathcal{S}'$.[6]
— Let A be a propositional attitude. A is *preserved under substructures* iff for any plausibility orders $\mathcal{S}$, $\mathcal{S}'$ and proposition P: if $\mathcal{S} \models AP$ and $\mathcal{S}'$ is a substructure of $\mathcal{S}$, then $\mathcal{S}' \models AP$.
— A is *persistent* if for any plausibility orders $\mathcal{S}$, $\mathcal{S}'$ and proposition P: if $\mathcal{S} \models AP$ and $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, then $\mathcal{S}' \models AP$.

The main intuition is that propositional attitudes that are preserved under substructures capture *purely universal* properties of a given plausibility order that do not depend on the presence or absence of particular worlds in a given structure; persistent propositional attitudes, on the other hand, capture properties that are *absolutely stable*, no matter what new information may be received from any source, regardless of the level of trust or distrust placed in the source. Since they are absolutley stable in this sense, persistent propositional attitudes are also preserved under substructures.

[6]Recall that $\mathcal{S}|_Q$ is the conditionalization of $\mathcal{S}$ on Q, given by $\mathcal{S}|_P := (S \cap P, \{(w,v) \in \mathcal{S} \mid w,v \in P\})$, cf. §1.1.3.

4.2.1. Examples. Knowledge K is preserved under substructures and, indeed, persistent. The same goes for triviality $\top$ and absurdity $\bot$. Another persistent propositional attitude is the disjunction of knowledge and the opposite of knowledge: $K \vee K^{\neg}$.

Two propositional attitudes that are preserved under substructures but *not* persistent are refinedness R and the disjunction of strong belief and opposite knowledge $Sb \vee K^{\neg}$ (the fixed point of strong trust $\Uparrow$).[7]

To see that refinedness R is preserved under substructures, an informal remark should suffice: if, for some proposition P, there are no ties between P-worlds and non-P-worlds in a given order $\mathcal{S}$, then this property will extend to substructures of $\mathcal{S}$. As an example showing that refinedness R is not persistent, consider the two orders $\mathcal{S} = (\{w, v\}, \{(w, v), (w, w), (v, v)\})$ and $\mathcal{S}' = (\{w, v\}, \{(w, v), (v, w), (w, w), (v, v)\})$. We notice that $\mathcal{S} \vDash R\{w\}$, $\mathcal{S} \twoheadrightarrow \mathcal{S}'$, but $\mathcal{S}' \nvDash R\{w\}$, because in $\mathcal{S}'$, w is tied with v, i.e., $w \approx'_{\mathcal{S}} v$. So refinedness R is not persistent.[8]

As for $Sb \vee K^{\neg}$: suppose a given order $\mathcal{S}$ satisfies $SbP \vee K^{\neg}P$. Assuming that $\mathcal{S}$ satisfies $K^{\neg}P$ (which just means that $\mathrm{S} \subseteq \neg P$), the same will obviously hold for any substructure of $\mathcal{S}$; and assuming that $\mathcal{S}$ satisfies SbP (which just means that all P-worlds are better than all non-P-worlds), any substructure will obviously satisfy either SbP or $K^{\neg}P$. So $Sb \vee K^{\neg}$ is preserved under substructures. However, $Sb \vee K^{\neg}$ is not preserved under refinements. For a counterexample, consider again the two orders $\mathcal{S} = (\{w, v\}, \{(w, v), (w, w), (v, v)\})$ and $\mathcal{S}' = (\{w, v\}, \{(w, v), (v, w), (w, w), (v, v)\})$. We notice that $\mathcal{S} \vDash Sb\{w\}$, hence $\mathcal{S} \vDash Sb\{w\} \vee K^{\neg}\{w\}$. Also, $\mathcal{S} \twoheadrightarrow \mathcal{S}'$. But $\mathcal{S}' \nvDash Sb\{w\}$, and also $\mathcal{S}' \nvDash K^{\neg}\{w\}$, so $\mathcal{S}' \nvDash Sb\{w\} vee e K^{\neg}\{w\}$. So $Sb \vee K^{\neg}$ is not persistent.

4.2.2. Distributivity. In line with the above remarks, our main interest is to ask: what properties of dynamic attitudes guarantee fixed points that are, respectively, preserved under substructures, or persistent? Answering this question is obviously key to arriving at a dynamic characterization of the propositional attitudes that are, respectively, preserved under substructures, or persistent. The starting point of our analysis is the notion of *distributivity*, defined next.

[7] Recall the definition of refinedness R from §2.5.1: $\mathcal{S} \vDash RP$ iff $\forall w, v \in \mathrm{S} : w \in P, v \notin P \Rightarrow w \not\approx v$. This means that RP is satisfied in an order $\mathcal{S}$ iff there are no ties between P-worlds and non-P-worlds. The attitude R should not be confused with refined belief Rb, as defined in §1.2.8.

[8] Using a similar argument, one can show that refined belief Rb (cf. §1.2.8 for the definition) is not persistent. But notice that refined belief is not even preserved under substructures!

A dynamic attitude τ is *distributive* iff

$$\mathcal{S}^{\tau P}|_Q = (\mathcal{S}|_Q)^{\tau P}$$

for any order $\mathcal{S}$ and propositions P and Q.[9]

Distributivity of a dynamic attitude τ means, roughly, that upgrades τP *commute* with arbitrary restrictions: applying an upgrade τP to an order $\mathcal{S}$ and then restricting the resulting order to Q yields the same result as restricting $\mathcal{S}$ to Q first and *then* applying τP.

In particular, this means that we can recover $\mathcal{S}^{\tau P}$ from applying τP to the individual pairs of worlds contained in $\mathcal{S}$. In fact, distributivity can equivalently be defined in this way. This is spelled out in the following lemma.

LEMMA 71. *Let τ be a dynamic attitude. The following are equivalent:*

1. *τ is distributive.*
2. *For any plausibility order $\mathcal{S}$, proposition P and worlds $w, v \in \mathrm{S}^{\tau P}$: $(w, v) \in \mathcal{S}^{\tau P}$ iff $(w, v) \in (\mathcal{S}|_{\{w,v\}})^{\tau P}$.*

PROOF. From (1.) to (2.), suppose that τ is distributive. Let $\mathcal{S}$ be a plausibility order, and let P be a proposition. Let $w, v \in \mathrm{S}^{\tau P}$. Then $(w, v) \in (\mathcal{S}|_{\{w,v\}})^{\tau P}$ iff (by distributivity) $(w, v) \in \mathcal{S}^{\tau P}|_{\{w,v\}}$ iff $(w, v) \in \mathcal{S}^{\tau P}$. This finishes the direction from (1.) to (2.).

From (2.) to (1.), suppose that for any plausibility order $\mathcal{S}$, proposition P and worlds $w, v \in \mathrm{S}^{\tau P}$: $(w, v) \in \mathcal{S}^{\tau P}$ iff $(w, v) \in (\mathcal{S}|_{\{w,v\}})^{\tau P}$. Let $\mathcal{S}$ be a plausibility order and P and Q propositions. We have to show that $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}^{\tau P}|_Q$. For one half of this, suppose that $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. We notice that this implies that $w, v \in \mathrm{S} \cap Q$. By the assumption, $(w, v) \in (\mathcal{S}|_{\{w,v\}})^{\tau P}$. Again using the assumption, $(w, v) \in \mathcal{S}^{\tau P}$. Since $w, v \in \mathrm{S} \cap Q$, it follows that $(w, v) \in \mathcal{S}^{\tau P}|_Q$. This shows one half of the equality. For the other half, suppose that $(w, v) \in \mathcal{S}^{\tau P}|_Q$. We notice again that this implies that $w, v \in \mathrm{S} \cap Q$. Now since $(w, v) \in \mathcal{S}^{\tau P}|_Q$, it follows that $(w, v) \in \mathcal{S}^{\tau P}$. By the assumption, $(w, v) \in (\mathcal{S}|_{\{w,v\}})^{\tau P}$. Since $w, v \in \mathrm{S} \cap Q$, again by the assumption, it follows that $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. This shows the other half. We conclude that $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}^{\tau P}|_Q$, and this yields the desired result: τ is distributive. This finishes the direction from (2.) to (1.). ⊣

So distributivity is characterized by a specific form of "acontextuality": one may recover the result of the upgrade τP in $\mathcal{S}$ by considering all "pair

[9]The reader may wonder if "distributivity" is an appropriate name for this property. We motivate this choice of terminology in §4.3 below, where we consider the historical roots of the notion in the dynamic semantics literature.

orders" $\mathcal{S}|_{\{w,v\}}$ living inside of $\mathcal{S}$ *in isolation*, applying τP to them, and taking the union of the all orders obtained in this way. This makes it clear what we mean by "acontextuality": to determine the relative plausibility of two worlds w and v in the new order $\mathcal{S}^{\tau P}$, all the information we need is given by the relative plausibility of w and v in the old order $\mathcal{S}$.

4.2.3. Examples. As the reader may want to check, of the examples of dynamic attitudes discussed so far, infallible (dis)trust $!^{(\neg)}$, strong (dis)trust $\Uparrow^{(\neg)}$, neutrality *id* and isolation $\varnothing$ are distributive. Minimal trust $\uparrow$, on the other hand, is an example of a non-distributive dynamic attitude.

Observe that distributivity is not preserved under strictures: while strong trust $\Uparrow$ is distributive, strong *positive* trust $\Uparrow^{+}$ is not. Intuitively, the reason is that from looking at a pair $(w, v) \in \mathcal{S}$ such that $w, v \notin P$ in isolation, one is not "able to tell" whether $w, v \in \mathcal{S}^{\Uparrow^{+}P}$: it depends on whether P-worlds are to be found somewhere in $\mathcal{S}$. This violates the "acontextuality" feature identified above.

We notice that the fixed points of distributive dynamic attitudes are preserved under substructures:

proposition 72. *Let τ be a dynamic attitude. If τ is distributive, then $\overline{\tau}$ is preserved under substructures.*

proof. Let τ be a distributive attitude. Consider an order $\mathcal{S}$ and proposition P such that $\mathcal{S} \vDash \overline{\tau}P$, i.e., $\mathcal{S}^{\tau P} = \mathcal{S}$. Pick an arbitrary $Q \subseteq W$. From our assumption it follows that $\mathcal{S}^{\tau P}|_Q = \mathcal{S}|_Q$. By distributivity, $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}^{\tau P}|_Q$. Hence $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}|_Q$. Thus $\mathcal{S}|_Q \vDash \overline{\tau}P$. So $\overline{\tau}$ is preserved under substructures. ⊣

So distributive attitudes fall squarely under the current topic of consideration. It turns out, however, that distributivity is slightly *too strong* a property to be useful in obtaining a characterization of the propositional attitudes that are preserved under substructures. That is: there are propositional attitudes that are preserved under substructures, but which are *not* the fixed point of any distributive dynamic attitude. An example of this phenomenon is the disjunction of knowledge and opposite knowledge, $K \vee K^{\neg}$. As pointed out above, $K \vee K^{\neg}$ is preserved under substructures. However:

proposition 73. *There exists no distributive dynamic attitude τ such that $\overline{\tau} = K \vee K^{\neg}$.*

proof. Let τ be a dynamic attitude. Suppose that the fixed point of τ is $K \vee K^{\neg}$. Towards a contradiction, suppose that τ is distributive. Consider an order $\mathcal{S} = (\{x, y\}, \leq_{\mathcal{S}})$ such that $x \in P$, $y \notin P$. Clearly, $\mathcal{S}|_{\{x\}} \vDash \overline{\tau}P$, $\mathcal{S}|_{\{y\}} \vDash \overline{\tau}P$. By

distributivity, $y, z \in \mathcal{S}^{\tau P}$, i.e., $S^{\tau P} = \{x, y\}$. We also know that $\mathcal{S}^{\tau P} \vDash \overline{\tau} P$ (since τ is idempotent). On the other hand, from the assumption that $\overline{\tau} = K \vee K^{\neg}$, together with the fact that $x \in P$, $y \notin P$, we conclude that $\mathcal{S}^{\tau P} \nvDash \overline{\tau} P$. This is a contradiction. So τ is not distributive. ⊣

Working towards a characterization of the propositional attitudes that *are* preserved under substructures, we shall thus require a weaker notion. It turns out that working with only *one half* of distributivity serves our purpose.

4.2.4. Semi-Distributivity. A dynamic attitude τ is *semi-distributive* if for any plausibility order $\mathcal{S}$ and proposition P and Q, we have: if $(w, v) \in \mathcal{S}^{\tau P}|_Q$, then $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$.

Notice that this property addresses the above counter-example: there exists a semi-distributive dynamic attitude whose fixed point is $K \vee K^{\neg}$. For example, the test $?K \vee K^{\neg}$ is semi-distributive.

4.2.5. Three Characterization Results. We are ready to state, prove and discuss our main results. Our first result characterizes the propositional attitudes that are preserved under substructures in terms of semi-distributivity and one additional property, *restrictiveness*.

A dynamic attitude τ is *restrictive* if for any order $\mathcal{S}$ and proposition P, there exists a proposition Q such that $\mathcal{S}^{\tau P} = \mathcal{S}|_Q$.

A dynamic attitude τ is thus restrictive if it does not affect the relative hierarchy between worlds in a given plausibility order, other than by outright *deleting* some of them. So for any order $\mathcal{S}$: $\mathcal{S}^{\tau P}$ can be obtained by restricting $\mathcal{S}$ to some proposition Q.

theorem 74. *Let A be a propositional attitude. The following are equivalent:*

1. *A is preserved under substructures.*

2. *There exists a semi-distributive, restrictive attitude τ such that $\overline{\tau} = A$.*

proof. From (1.) to (2.), let A be an introspective propositional attitude that is preserved under submodels. Recall that $?A$ is the dynamic attitude given by

$$\mathcal{S}^{?AP} := \begin{cases} \mathcal{S} & \mathcal{S} \vDash AP, \\ \varnothing & \text{otherwise.} \end{cases}$$

The fixed point of $?A$ is A; also $?A$ is restrictive, since for any order $\mathcal{S}$ and proposition P, either $\mathcal{S}^{?AP} = \mathcal{S}|_S$ or $\mathcal{S}^{?AP} = \mathcal{S}|_\varnothing$. To complete the proof of this direction, it is thus sufficient to show that $?A$ is semi-distributive. Suppose,

then, that for some $Q \subseteq W$, we have $(w, v) \in \mathcal{S}^{?AP}|_Q$. Then $\mathcal{S}^{?AP}|_Q \neq \varnothing$ (in particular: $w, v \in \mathrm{S}^{?AP} \cap Q$), so $\mathcal{S}^{?AP} = \mathcal{S}$, which implies that $\mathcal{S} \vDash AP$. Since A is preserved under substructures, also $\mathcal{S}|_Q \vDash AP$. Since the fixed point of $?A$ is A, $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}|_Q$. Since $w, v \in \mathrm{S}$ and $w, v \in Q$, it follows that $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. So $?A$ is semi-distributive, which completes the "(1.) to (2.)" direction.

From (2.) to (1.), suppose that $\mathcal{S} \vDash \overline{\tau} P$. We have to show that $\mathcal{S}|_Q \vDash \overline{\tau} P$, which is to say that $(\mathcal{S}|_Q)^{\tau P} = \mathcal{S}|_Q$. One half of the equality: from the fact that $(w, v) \in \mathcal{S}|_Q$, we conclude that $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$ by semi-distributivity. The other half of the equality: suppose that $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. By restrictiveness, $(\mathcal{S}|_Q)^{\tau P} = (\mathcal{S}|_Q)|_{Q'}$ for some $Q' \subseteq W$. And since $w, v \in Q$, we conclude that $(w, v) \in \mathcal{S}|_Q$, which completes the other half, and the proof. ⊣

This characterization will prove useful below. Still, the result leaves something to be desired. Consider strong trust $\Uparrow$. The fixed point of $\Uparrow$ is $Sb \vee K^{\neg}$, the disjunction of strong belief and opposite knowledge. As pointed out earlier, $Sb \vee K^{\neg}$ is preserved under substructures. However, $\Uparrow$, while distributive (and thus semi-distributive), is *not* restrictive: it does not hold for all plausibility orders $\mathcal{S}$ that $\mathcal{S}^{\Uparrow P}$ is the restriction of $\mathcal{S}$ to some proposition Q. The reason is simple: quite often, applying $\Uparrow P$ leads to changes in the relative hierarchy of worlds in a given order $\mathcal{S}$.

So $\Uparrow$ lies, as it were, outside of the "scope" of the theorem. Note that this is not a counterexample against the result: the result merely claims the *existence* of a dynamic attitude with the desired properties; there is no claim to the extent that *any* dynamic attitude whose fixed point is preserved under substructures is semi-distributive and restrictive. Still, there is room for improvement.

We call a dynamic attitude τ *discerning* if for any plausibility order $\mathcal{S}$, proposition P and worlds $w, v \in \mathrm{S}^{\tau P}$, we have: if $w \approx_{\tau P} v$, then $w \approx v$.

A dynamic attitude τ is thus discerning if applying it to a plausibility order $\mathcal{S}$ *does not introduce ties*: worlds that are equiplausible in $\mathcal{S}^{\tau P}$ have already been equiplausible in $\mathcal{S}$.

Observe that restrictive dynamic attitudes are discerning (they don't "equalize" pairs of worlds, they merely kill worlds), but not the other way around (counterexample: $\Uparrow$ is discerning, but not restrictive). From the perspective of the above considerations, the following characterization is improved:

THEOREM 75. *Let A be an introspective propositional attitude. The following are equivalent:*

1. *A is preserved under substructures.*

2. *There exists a discerning, semi-distributive dynamic attitude τ such that $\overline{\tau} = A$.*

PROOF. The direction from (1.) to (2.) is analogous to the previous theorem: we work with tests again, noticing that $?A$ is discerning.

From (2.) to (1.), let τ be discerning and semi-distributive. We have to show that the fixed point of τ is preserved under substructures. So suppose that $\mathcal{S} \vDash \overline{\tau}P$. We need to establish that $\mathcal{S}|_Q \vDash \overline{\tau}P$, i.e.: that $\mathcal{S}|_Q = (\mathcal{S}|_Q)^{\tau P}$. For one half of this, suppose that $(w, v) \in \mathcal{S}|_Q$. Then $(w, v) \in \mathcal{S}$, and by the assumption, $(w, v) \in \mathcal{S}^{\tau P}$, so $(w, v) \in \mathcal{S}^{\tau P}|_Q$, and since τ is semi-distributive, $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. For the other half, suppose $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$. Towards a contradiction, suppose that $(w, v) \notin \mathcal{S}|_Q$. This entails that $(w, v) \notin \mathcal{S}$, so $(v, w) \in \mathcal{S}$ (by totality of plausibility orders), i.e., $v <_{\mathcal{S}} w$, hence $v <_{\mathcal{S}|_Q} w$ (since $w, v \in Q$). Now from the fact that $(v, w) \in \mathcal{S}$ together with our assumption that $\mathcal{S} \vDash \overline{\tau}P$, we conclude that $(v, w) \in \mathcal{S}^{\tau P}$. Since τ is semi-distributive, this entails that $(v, w) \in (\mathcal{S}|_Q)^{\tau P}$. Since also $(w, v) \in (\mathcal{S}|_Q)^{\tau P}$—our assumption—, we have $w \approx_{(\mathcal{S}|_Q)^{\tau P}} v$. Since τ is discerning, $w \approx_{\mathcal{S}|_Q} v$. But as we have seen above, $v <_{\mathcal{S}|_Q} w$. This is a contradiction. It follows that $(w, v) \in \mathcal{S}_Q$, after all. So $\mathcal{S}|_Q \vDash \overline{\tau}P$, and the second half of the "(2.) to (1.)" direction is complete. ⊣

Note that strong trust $\Uparrow$ is semi-distributive and discerning, hence, the fact that its fixed point is preserved under substructures follows directly from the above result.

To characterize the *persistent* propositional attitudes, we add an additional property to the mix: *domain-stability*.

A dynamic attitude τ is domain-stable if for any plausibility orders $\mathcal{S}$ and $\mathcal{T}$: if $\mathrm{S} = \mathrm{T}$, then $\mathrm{S}^{\tau P} = \mathrm{T}^{\tau P}$.

Domain-stability expresses that the new domain we obtain when applying an upgrade to a plausbility order $\mathcal{S}$ does not depend on the relative plausibility of the worlds in $\mathcal{S}$: any plausibility order $\mathcal{T}$ based on the same domain as $\mathcal{S}$ yields the same new domain when the upgrade is applied (while $\leq_{\mathcal{S}^{\tau P}}$ and $\leq_{\mathcal{T}^{\tau P}}$ may, of course, differ).

THEOREM 76. *Let A be a propositional attitude. The following are equivalent:*

1. *A is persistent.*

2. *There exists a semi-distributive, restrictive, domain-stable dynamic attitude τ such that $\overline{\tau} = A$.*

PROOF. From (1.) to (2.), suppose that A is persistent. We consider the test for A, $?A$. As for semi-distributivity and restrictiveness: since A is persistent, A is preserved under substructures. We can thus argue as in the proof of Theorem 74 to conclude that $?A$ is semi-distributive and restrictive. To prove our claim, it remains to show that $?A$ is domain-stable. Let $\mathcal{S}$ be a plausibility

order and P a proposition. Let $\mathcal{T}$ be an order-variant of $\mathcal{S}$. We have to show that $\mathrm{T}^{?AP} = \mathrm{S}^{?AP}$. We consider two cases. First, suppose that $\mathcal{S} \models AP$. Then $\mathcal{S}^{?AP} = \mathcal{S}$. Since A is persistent, $\mathcal{T} \models AP$. So $\mathcal{T}^{?AP} = \mathcal{T}$. Since $\mathcal{T}$ and $\mathcal{S}$ are order-variants, it follows that $\mathrm{S}^{?AP} = \mathrm{T}^{?AP}$. The claim holds. Second, suppose that $\mathcal{S} \not\models AP$. Then $\mathcal{S}^{?AP} = \varnothing$. Assuming that $\mathcal{T} \models AP$, we derive a contradiction, observing that $\mathcal{T} \twoheadrightarrow \mathcal{S}$, so $\mathcal{S} \models AP$ by the fact that A is persistent. Hence $\mathcal{T} \not\models AP$. So $\mathcal{T}^{?AP} = \varnothing$. But then, $\mathrm{S}^{?AP} = \varnothing = \mathrm{T}^{?AP}$: again, the claim holds. So $?A$ is domain-stable. We have thus shown that there exists a semi-distributive, restrictive and domain-stable dynamic attitude τ such that $\overline{\tau} = A$, and this finishes one direction.

From (2.) to (1.), suppose that there exists a semi-distributive, restrictive, domain-stable dynamic attitude τ such that $\overline{\tau} = A$. We have to show that A is persistent, which by our assumption amounts to proving that $\overline{\tau}$ is persistent. So let $\mathcal{S}$ be a plausibility order, P a proposition, suppose that $\mathcal{S} \models \overline{\tau}P$, and let $\mathcal{T}$ be a plausibility order such that $\mathcal{S} \twoheadrightarrow \mathcal{T}$. We have to show that $\mathcal{T} \models \overline{\tau}P$.

Take any substructure $\mathcal{U}$ of $\mathcal{S}$ with the same domain as $\mathcal{T}$, i.e., consider $\mathcal{U}$ such that $\mathrm{U} = \mathrm{T}$. Since τ is semi-distributive and restrictive, we apply Theorem 74 to conclude that $\mathcal{U} \models \overline{\tau}P$. Since τ is domain-stable, it follows that $\mathrm{T}^{\tau P} = \mathrm{T}$. Since τ is restrictive, we conclude that $\mathcal{T}^{\tau P} = \mathcal{T}|_{\mathrm{T}}$. But this just says that $\mathcal{T}^{\tau P} = \mathcal{T}$. Thus $\overline{\tau} = A$ is persistent. This finishes the second direction, and the proof. ⊣

Theorem 74, Theorem 75 and 76 provide just a sample of what we feel could be done in the area. It would, for example, be interesting to have similar characterizations of the propositional attitudes that are preserved under *refinements* (cf. §3.6.1); example: refinedness R is preserved under refinements. Also, the propositional attitudes that are preserved under arbitrary *reorderings* of worlds are of interest; knowledge K is of this kind.[10] Going the other direction, one would like to know if there is a natural preservation property capturing *semi-distributivity*, which has played an important role in the analysis of this chapter. We leave these questions for future research.

The remainder of this chapter is devoted to a study of the property that was the initial starting point of our analysis in this section: distributivity. We are thus focusing on a particular *subclass* of the dynamic attitudes that are preserved under substructures. Our main result is a tight characterization of the upgrades given by positive distributive dynamic attitudes. We start by clarifying the roots of the notion of distributivity.

[10] A propositional attitude A is *preserved under reorderings* if for any plausibility orders $\mathcal{S}$, $\mathcal{S}'$ and proposition P: if $\mathrm{S} = \mathrm{S}'$ and $\mathcal{S} \models AP$, then $\mathcal{S}' \models AP$. That is: the particular hierarchy on the worlds in $\mathcal{S}$ imposed by the relation $\leq$ does not matter, as far as satisfaction of AP is concerned.

4.3. *Distributivity in Dynamic Semantics*

The notion of distributivity has its roots in a closely related notion of the same name that has been prominent in discussions in the dynamic semantics literature. The purpose of this section is to clarify the connection.

Let W be a set of possible worlds. According to a familiar picture, widespread in dynamic semantics, information states can be captured by propositions, and types of changes that may occur can be captured by functions on $\wp(W)$.[11] For the purposes of this section, we shall refer to functions on $\wp(W)$ as *change potentials*.

A change potential u is *distributive* iff

$$(S \cup T)^u = S^u \cup T^u,$$

and u is *eliminative* iff

$$S^u \subseteq S$$

for any $S \subseteq W$.

The much-cited next observation is due to van Benthem (1986):[12]

THEOREM 77 (van Benthem (1986)). *Let u be a change potential. The following are equivalent:*

1. *u is eliminative and distributive.*

2. *$S^u = S \cap W^u$ for all $S \subseteq W$.*

PROOF. From (1.) to (2.), start with $S \cap W^u = S \cap (S \cup \neg S)^u$. By distributivity, this is the same as $S \cap (S^u \cup (\neg S)^u)$. By properties of sets, this equals $(S \cap S^u) \cup (S \cap (\neg S)^u)$. By eliminativeness, the first disjunction equals S^u, and the second disjunct equals $\varnothing$. But $S^u \cup \varnothing = S^u$. So $S \cap W^u = S^u$.

From (2.) to (1.), suppose that $S^u = S \cap W^u$ for all $S \subseteq W$. It is immediate that u is eliminative. Observe also that, by our assumption, given propositions S and T, we have $S^u \cup T^u = (S \cap W^u) \cup (T \cap W^u) = (S \cup T) \cap W^u = (S \cup T)^u$. So u is distributive. ⊣

The collection of all distributive and eliminative change potentials over $\wp(W)$ is thus given by

$$\{S \mapsto S \cap P\}_{P \subseteq W}.$$

[11] Compared to the setting of this dissertation, propositions seen as information states adopt the role taken by plausibility orders, and functions on $\wp(W)$ adopt the role taken by upgrades.

[12] Cf. also van Eijck and Visser (2008) and Rothschild and Yalcin (2012) for discussion.

What this means is that any distributive and eliminative change potential u can, essentially, be *identified* with the proposition W^u. This observation is often taken to indicate that distributive and eliminative observations really do not have to offer anything that goes beyond (a static conception of) propositional content: the type of change captured by a distributive and eliminative change potential may just as well be given by a *single* proposition with which given information states are intersected. An example of a non-distributive change potential is $!^\sim P$ (i.e., bare semi-trust $!^\sim$ applied to P, cf.§2.5.3), given by

$$S^{!^\sim P} := \begin{cases} S & S \cap P \neq \varnothing \\ \varnothing & \text{otherwise} \end{cases}$$

Turning to our main thread of discussion: how does the notion of distributivity (of a dynamic attitude) we have introduced in §4.2 relate to the notion of distributivity for change potentials defined above? The notion of distributivity is not immediately transferable to plausibility orders, because the union of two total preorders is in general neither total nor transitive. However, we notice:

PROPOSITION 78. *Let u be an eliminative change potential over $\wp(W)$. Then u is distributive iff for any propositions S and Q:*

$$S^u \cap Q = (S \cap Q)^u.$$

PROOF. Let u be an eliminative change potential over $\wp(W)$. Suppose u is distributive. Then $S^u \cap Q = (S \cap W^u) \cap Q = S \cap Q \cap W^u = (S \cap Q)^u$. This completes one half.

For the other half, suppose that for any S: $S^u \cap Q = (S \cap Q)^u$. We have to show that u is distributive. Let $S, T \subseteq W$. Notice that $S^u = ((S \cup T) \cap S)^u = (S \cup T)^u \cap S$ by our assumption, and analogously, $T^u = (S \cup T)^u \cap T$. So $S^u \cup T^u = ((S \cup T)^u \cap S) \cup ((S \cup T)^u \cap T)$. The latter is the same as $(S \cup T)^u \cap (S \cup T)$. By eliminativeness, this is $(S \cup T)^u$. So $S^u \cup T^u = (S \cup T)^u$, the desired result: u is distributive. ⊣

Putting $P|_Q := P \cap Q$, the previous proposition says that a change potential u is distributive iff for any proposition S and Q: $S^u|_Q = (S|_Q)^u$. This motivates our definition of distributivity for dynamic attitudes, according to which a dynamic attitude τ is distributive iff for any order $\mathcal{S}$ and propositions P and Q: $(\mathcal{S}^{\tau P})|_Q = (\mathcal{S}|_Q)^{\tau P}$, i.e., this is simply the natural analogue for plausibility orders.[13]

[13]Notice that the notion of eliminativeness which is used in the proof of Proposition 78 is, in a sense, automatically satisfied by an upgrade u, since we require that for any plausibility order $\mathcal{S}$: $S^u \subseteq S$.

4.4. Positive Distributive Dynamic Attitudes

Distributive dynamic attitudes impose a lot of structure on the upgrades they give rise to. We begin this section by establishing that distributive dynamic attitudes enjoy the properties of *selectiveness* and *conservation*. This allows us to show that the transformations given by *positive* distributive attitudes are actually of exactly two possible types: they are lexicographic upgrades or updates (i.e., they can be described by upgrades of the form $\Uparrow P$ or $!P$). As a corollary, we obtain characterizations of strong trust and infallible trust.

4.4.1. Selectiveness. A dynamic attitude τ is *selective* if for any plausibility order $\mathcal{S}$ and proposition P: $S^{\tau P} \in \{S, \varnothing, S \cap P, S \cap \neg P\}$.

lemma 79. *Distributive dynamic attitudes are selective.*

proof. Let τ be a distributive dynamic attitude, let $\mathcal{S}$ be a plausibility order, and P a proposition. We have to show that $S^{\tau P} \in \{S, \varnothing, S \cap P, S \cap \neg P\}$. Consider the plausibility orders $\mathcal{S}^{\tau P}|_P$ and $\mathcal{S}^{\tau P}|_{\neg P}$. Since τ is distributive, $\mathcal{S}^{\tau P}|_P = (\mathcal{S}|_P)^{\tau P}$ and $\mathcal{S}^{\tau P}|_{\neg P} = (\mathcal{S}|_{\neg P})^{\tau P}$ (call these two equalities (†)). By strong informativity, $(\mathcal{S}|_P)^{\tau P} \in \{\mathcal{S}|_P, \varnothing\}$ and $(\mathcal{S}|_{\neg P})^{\tau P} \in \{\mathcal{S}|_{\neg P}, \varnothing\}$. By (†), we conclude that $\mathcal{S}^{\tau P}|_P \in \{\mathcal{S}|_P, \varnothing\}$ and $\mathcal{S}^{\tau P}|_{\neg P} \in \{\mathcal{S}|_{\neg P}, \varnothing\}$. So $S^{\tau P} \cap P \in \{S \cap P, \varnothing\}$ and $S^{\tau P} \cap \neg P \in \{S \cap \neg P, \varnothing\}$. Since, by basic set-theory, $S^{\tau P} = (S^{\tau P} \cap P) \cup (S^{\tau P} \cap \neg P)$, it follows that $S^{\tau P} \in \{(S \cap P) \cup (S \cap \neg P), (S \cap P) \cup \varnothing, (S \cap \neg P) \cup \varnothing, \varnothing \cup \varnothing\}$. Which is to say: $S^{\tau P} \in \{S, S \cap P, S \cap \neg P, \varnothing\}$. But this is exactly our claim. ⊣

4.4.2. Conservation. Recall from §3.7.1 that a dynamic attitude τ is *conserving* iff for any plausibility order $\mathcal{S}$ and proposition P, and for any $w, v \in S^{\tau P}$: if $w \in P$ iff $v \in P$, then $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}} v$.

lemma 80. *Distributive dynamic attitudes are conserving.*

proof. Let τ be a distributive attitude, let $\mathcal{S}$ be a plausibility order, let P be a proposition, let $w, v \in S^{\tau P}$, and suppose that $w \in P$ iff $v \in P$. We have to show that $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}} v$.

Consider the order $\mathcal{S}|_{\{w,v\}}$. Since $w \in P$ iff $v \in P$, it follows that $\{w, v\} \cap P \in \{\{w, v\}, \varnothing\}$. It follows by strong informativity that $(\mathcal{S}|_{\{w,v\}})^{\tau P} \in \{\mathcal{S}|_{\{w,v\}}, \varnothing\}$. By distributivity, $(\mathcal{S}|_{\{w,v\}})^{\tau P} = \mathcal{S}^{\tau P}|_{\{w,v\}}$. So $\mathcal{S}^{\tau P}|_{\{w,v\}} \in \{\mathcal{S}|_{\{w,v\}}, \varnothing\}$. But $w, v \in S^{\tau P}$, so $\mathcal{S}^{\tau P}|_{\{w,v\}} \neq \varnothing$. Thus $\mathcal{S}^{\tau P}|_{\{w,v\}} = \mathcal{S}|_{\{w,v\}}$. Hence $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}} v$. ⊣

4.4.3. Positive Distributive Dynamic Attitudes. The previous results allow us to show, as announced, that the transformations given by *positive* distributive dynamic attitudes are generally of two types only: they are lexicographic upgrades or updates.

proposition 81. *Let τ be a positive distributive dynamic attitude. Then for any plausibility order $\mathcal{S}$ and proposition P: $\mathcal{S}^{\tau P} \in \{\mathcal{S}^{!P}, \mathcal{S}^{\Uparrow P}\}$.*

proof. Let $\mathcal{S}$ be a plausibility order, and P a proposition. We discuss two cases. First, suppose that P is insubstantial in $\mathcal{S}$, i.e., either $P \cap S = \varnothing$ or $P \cap S = S$. Then it follows that $\mathcal{S}^{\tau P} \in \{\mathcal{S}, \varnothing\}$ by strong informativity (Lemma 6). If $\mathcal{S}^{\tau P} = \varnothing$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{!P}$. If, on the other hand, $\mathcal{S}^{\tau P} = \mathcal{S}$, then $\mathcal{S}^{\tau P} = \mathcal{S}^{\Uparrow P}$. Either way, the claim holds.

Now assume that P is substantial in $\mathcal{S}$. By Lemma 79, $S^{\tau P} \in \{S, \varnothing, S \cap P, S \cap \neg P\}$. Since τ is positive, $S^{\tau P} \neq \varnothing$. So $S^{\tau P} \in \{S, \varnothing, S \cap P, S \cap \neg P\}$. Since τ is positive, $S^{\tau P} \cap P \neq \varnothing$, using the fact that P is substantial in $\mathcal{S}$. So $S^{\tau P} \neq S \cap \neg P$. So $S^{\tau P} \in \{S, S \cap P\}$.

We thus have two sub-cases to consider. Suppose first that $S^{\tau P} = S$. Let $w, v \in S$ such that $w \in P$, $v \notin P$. Since τ is positive, $w <_{(\mathcal{S}|_{\{w,v\}})^{\tau P}} v$. Since τ is distributive, $w <_{\mathcal{S}^{\tau P}|_{\{w,v\}}} v$, So $w <_{\mathcal{S}^{\tau P}} v$. Now let $w, v \in S$ such that $w \in P$ iff $v \notin P$. By Lemma 80, $w \leq_{\mathcal{S}} v$ iff $w \leq_{\mathcal{S}^{\tau P}} v$. But this is exactly the definition of $\Uparrow$. So we may conclude that $\mathcal{S}^{\tau P} = \mathcal{S}^{\Uparrow P}$. The claim holds.

As the second sub-case, suppose that $S^{\tau P} = S \cap P$. Let $w, v \in S^{\tau P}$. By Lemma 80, $w \leq_{\mathcal{S}^{\tau P}} v$ iff $w \leq_{\mathcal{S}} v$. But then, $\mathcal{S}^{\tau P} = \mathcal{S}^{!P}$. Again, the claim holds, and this concludes the second case, and the proof. ⊣

As a corollary of the Proposition 81, we obtain characterizations of infallible trust and strong trust, using the following two properties:

— A dynamic attitude τ is *hard* iff for any plausibility order $\mathcal{S}$ and proposition P: $S^{\tau P} \subseteq S \cap P$.
— A dynamic attitude τ is *soft* iff for any plausibility order $\mathcal{S}$ and proposition P: $S^{\tau P} = S$.

If a dynamic attitude τ is hard, then, on receiving the information that P from a τ-source, the agent acquires hard information that P, i.e., all non-P-worlds are deleted; and if τ is soft, then, on receiving the information that P from a τ-source, the agent acquires *no* hard information.

corollary 82.

1. *Strong trust is the only dynamic attitude that is distributive, positive and soft.*

2. *Infallible trust is the only attitude that is distributive, positive and hard.*

PROOF.

1. It is easy to check that strong trust is distributive, positive and soft. As for the uniqueness claim, suppose that τ is distributive, positive and soft. Let $\mathcal{S}$ be a plausibility order, P a proposition. Assume first that $P \cap \mathrm{S} = \mathrm{S}$. Then $\mathcal{S}^{\Uparrow P} = \mathcal{S}^{!P}$. By Proposition 81, it follows that $\mathcal{S}^{\tau P} = \mathcal{S}^{\Uparrow P}$. Assume, second, that $P \cap \mathrm{S} \neq \mathrm{S}$. Then $\mathrm{S}^{!P} \neq \mathrm{S}$. But τ is soft, so $\mathrm{S}^{\tau P} = \mathrm{S}$, thus $\mathrm{S}^{\tau P} \neq \mathrm{S}^{!P}$. By Proposition 81, $\mathcal{S}^{\tau P} = \mathcal{S}^{\Uparrow P}$. So in either case, $\mathcal{S}^{\tau P} = \mathcal{S}^{\Uparrow P}$, and this shows that $\tau = \Uparrow$. So $\Uparrow$ is the only dynamic attitude that is distributive, positive and soft.

2. It is easy to check that infallible trust is distributive, positive and hard. As for the uniqueness claim, suppose that τ is distributive, positive and hard. Let $\mathcal{S}$ be a plausibility order, P a proposition. Assume first that $P \cap \mathrm{S} = \mathrm{S}$. Then $\mathcal{S}^{\Uparrow P} = \mathcal{S}^{!P}$. By Proposition 81, it follows that $\mathcal{S}^{\tau P} = \mathcal{S}^{!P}$. Assume, second, that $P \cap \mathrm{S} \neq \mathrm{S}$, which implies that $P \cap \mathrm{S} \subset \mathrm{S}$. Since τ is hard, we have that $\mathrm{S}^{\tau P} \subseteq P \cap \mathrm{S}$, and since $P \cap \mathrm{S} \subset \mathrm{S}$, it follows that $\mathrm{S}^{\tau P} \subset \mathrm{S}$. However, by definition of $\Uparrow$, $\mathrm{S}^{\Uparrow P} = \mathrm{S}$. So $\mathrm{S}^{\tau P} \neq \mathrm{S}^{\Uparrow P}$. Thus $\mathcal{S}^{\tau P} \neq \mathcal{S}^{\Uparrow P}$. By Proposition 81, $\mathcal{S}^{\tau P} = \mathcal{S}^{!P}$. So in either case, $\mathcal{S}^{\tau P} = \mathcal{S}^{!P}$, and this shows that $\tau = !$. So $!$ is the only dynamic attitude that is distributive, positive and hard. ⊣

This result motivates the claim that there is really only one difference between $\Uparrow$ and $!$: while the former is soft, the latter is hard. This confirms a common intuition that performing a lexicographic upgrade $\Uparrow P$ with some proposition P is "the same" as performing an update $!P$ with the same proposition, "just without deleting the (non-P-)worlds." We observe that the results of this section can easily be adapted to obtain analogous characterizations of strong *dis*trust $\Uparrow^{\neg}$ and infallible *dis*trust $!^{\neg}$.

Chapter 5. Logics

In this chapter, we move from a purely semantic to a properly logical setting, studying logical languages that contain operators for talking about dynamic attitudes in one way or another. We consider both the single-agent setting that we work with in most of this dissertation, and the multi-agent setting that we have introduced in §2.7.

The key construct of the single-agent language (studied in §§5.1–5.5) allows us to build sentences of the form

$$[s{:}\,\varphi]\psi$$

with the reading "after the agent receives the information that φ from a source of type s, ψ holds." Here, a source's being "of type s" is interpreted using the machinery developed in earlier chapters, i.e., we formalize the idea that to each type of source there corresponds a dynamic attitude capturing how the agent assesses the reliability of sources of that type.

The multi-agent language that we consider in this chapter (studied in §5.6) will allow us to build sentences of the form

$$[a{:}\,\varphi]\psi$$

with the reading "after the communication act $[a{:}\,\varphi]$, ψ holds." Here, the idea is that the agents use their dynamic attitude towards a to upgrade their plausibility order on an occasion where agent a asserts that φ. This will be made precise using the work of §2.7, where we have introduced a formal notion of communication act.

The key notion we work with is the notion of a definable dynamic attitude. Very roughly and generally speaking, a dynamic attitude is said to be definable in a modal language if the language supplies syntactic devices that are expressive enough to describe the effects of applying certain upgrades to the models in which the language is interpreted. In this chapter, we consider the more specific case of definability in the epistemic-doxastic language

which has operators for infallible and defeasible knowledge. The distinctive feature of the logics presented here compared to previous research is that they "work" not only for specific examples of belief revision policies, but for any dynamic attitude that is definable in the epistemic-doxastic language.

Introducing a logical setting with a proper syntax and semantics raises a number of familiar questions. We focus here on the two most basic, and most well-studied technical questions in dynamic epistemic logic, as they apply to our setting: expressivity and completeness.

5.1. The Epistemic-Doxastic Language

We have already seen the epistemic-doxastic language in §1.6. But since that has been a while ago, we develop all machinery from scratch.

5.1.1. Signatures. A *signature* is a tuple

$$(W, I, L, \Phi, \mathcal{A})$$

where

— W is a countably infinite set of possible worlds,
— I is a countable index set (called the set of *attitude labels*)
— L is a function assigning a dynamic attitude $\boldsymbol{\tau}$ to each attitude label $\tau \in I$,
— Φ is a set of symbols (called the *set of atomic sentences*), and
— $\mathcal{A}$ is a finite, non-empty set (called the *set of agents*)

The set $\mathcal{A}$ will only play a role for our discussion starting in §5.6, where we consider a multi-agent version of our setting. In the preceding sections, we will continue to discuss the single-agent setting familiar from most of the previous work in this dissertation, in which an agent receiving information, and the sources from which the agent receives this information, are merely implicit.

We shall refer, in this chapter, to attitude labels in I (which are to be thought of as "syntactic" objects) using lower-case greek letters (e.g., σ, τ), and to the corresponding dynamic attitudes (semantic objects) assigned to the labels using the corresponding bold-face lower-case greek letters (e.g., $\boldsymbol{\sigma}, \boldsymbol{\tau}$), as we have already done in the preceding definition. That is, given some attitude label τ, we shall always assume that $L(\tau)$ is given by $\boldsymbol{\tau}$.

Unless specifically noted otherwise, we assume an arbitrary but fixed signature $(W, I, L, \Phi, \mathcal{A})$ as given in the following.

5.1.2. Valuations. A *valuation* $[\![\cdot]\!]$ is a function

$$p \overset{[\![\cdot]\!]}{\longmapsto} [\![p]\!]$$

that assigns a proposition $P \subseteq W$ to each atomic sentence $p \in \Phi$.

Given a valuation $[\![\cdot]\!]$, an atomic sentence $p \in \Phi$ and a proposition $P \subseteq W$, we shall write $[\![\cdot]\!][p \mapsto P]$ for the valuation which is just like $[\![\cdot]\!]$, except that $[\![p]\!] = P$.

5.1.3. Plausibility Models. A *(single-agent) plausibility model* is a pair

$$\mathcal{M} = (\mathcal{S}, [\![\cdot]\!]),$$

where $\mathcal{S}$ is a plausibility order (on W), and $[\![\cdot]\!]$ is a valuation.

Given a plausibility model $\mathcal{M}$, an atomic sentence $p \in \Phi$ and a proposition $P \subseteq W$, we shall write $\mathcal{M}[p \mapsto P]$ for the plausibility model $(\mathcal{S}, [\![\cdot]\!]')$ which is just like $\mathcal{M}$ except that $[\![\cdot]\!]' = [\![\cdot]\!][p \mapsto P]$.

5.1.4. The Epistemic-Doxastic Language. The language $\mathcal{L}$ (called the *(single-agent) epistemic-doxastic language*) is given by the following grammar ($p \in \Phi$):

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid K\varphi$$

Read $K\varphi$ as *the agent infallibly (*or: *indefeasibly) knows that* φ; read $\Box\varphi$ as *the agent defeasibly knows that* φ.

We define $\top$ as $p \vee \neg p$ and set $\bot := \neg\top$.

5.1.5. Semantics. We interpret the language $\mathcal{L}$ in the usual manner, by providing, for each plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$, a map $[\![\cdot]\!]_{\mathcal{M}}$ that assigns a proposition $[\![\varphi]\!]_{\mathcal{M}} \subseteq S$ to each sentence $\varphi \in \mathcal{L}$, the proposition comprising the worlds where φ is *satisfied* in $\mathcal{M}$ (or: the worlds w such that φ is *true at* w in $\mathcal{M}$).

Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model. $[\![\cdot]\!]_{\mathcal{M}}$ is defined by induction on the construction of φ.

$$\begin{aligned}
[\![p]\!]_{\mathcal{M}} &:= [\![p]\!] \cap S, \\
[\![\neg\varphi]\!]_{\mathcal{M}} &:= S \setminus [\![\varphi]\!]_{\mathcal{M}}, \\
[\![\varphi \wedge \psi]\!]_{\mathcal{M}} &:= [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}}, \\
[\![\Box\varphi]\!]_{\mathcal{M}} &:= \Box_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}}, \\
[\![K\varphi]\!]_{\mathcal{M}} &:= K_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}},
\end{aligned}$$

where we recall the definitions of the propositional attitudes *defeasible knowledge* $\Box$ and *irrecovable knowledge* K (cf. §1.2.8), according to which

$$\Box_{\mathcal{S}}[\![\varphi]\!]_{\mathcal{M}} = \{w \in \mathrm{S} \mid \exists v \in \mathrm{S} : v \leq_{\mathcal{S}} w \text{ and } v \in [\![\varphi]\!]_{\mathcal{M}}\},$$

and

$$K_{\mathcal{S}}[\![\varphi]\!]_{\mathcal{M}} = \{w \in \mathrm{S} \mid \mathrm{S} \subseteq [\![\varphi]\!]_{\mathcal{M}}\}.$$

We shall use the notation $\mathcal{M}, w \vDash \varphi$ to mean that $w \in [\![\varphi]\!]_{\mathcal{M}}$. We say that a sentence $\varphi \in \mathcal{L}$ is *valid* iff $[\![\varphi]\!]_{\mathcal{M}} = \mathrm{S}$ for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$. We write $\vDash_{\mathcal{L}\varphi}$ if φ is valid.

5.1.6. Axiomatization. The epistemic-doxastic language $\mathcal{L}$ was axiomatized by Baltag and Smets (2008) using the following derivation system. The *logic of defeasible and indefeasible knowledge* L is given by following axioms and rules of inference:

— *Axioms:*
  — All instances of theorems of propositional calculus
  — $K(\varphi \to \psi) \to (K\varphi \to K\psi)$
  — $\Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$
  — The S5 axioms for K
  — The S4 axioms for $\Box$
  — $K\varphi \to \Box\varphi$
  — $K(\varphi \vee \Box\psi) \wedge K(\psi \vee \Box\varphi) \to (K\varphi \vee K\psi)$
— *Rules of inference:*
  — From φ and $\varphi \to \psi$ infer ψ
  — From φ infer $K\varphi$ and $\Box\varphi$

The fact that L is sound and complete w.r.t. plausibility models follows from the work of Baltag and Smets (2008). We argue as follows:

theorem 83 (Baltag and Smets (2008)). *The logic of defeasible and indefeasible knowledge* L *is weakly sound and complete w.r.t. plausibility models.*

proof. For soundness, observe that all the axioms are valid, and that the rules of inference preserve validity, and argue by induction on the length of a derivation in L. For completeness, let $\varphi \in \mathcal{L}$ be an L-consistent sentence (a sentence the negation of which is not provable in L). Baltag and Smets (2008) show (their Theorem 2.5) that there exists a triple $\mathcal{X} = (X, \leq, V)$, where X is a finite, non-empty set, $\leq$ a total preorder on X and V a function assigning a set

of worlds $Y \subseteq X$ to each atomic sentence $p \in \Phi$, and an element $x \in X$ such that $\mathcal{X}, x \vDash_{BS} \varphi$, with $\vDash_{BS}$ their truth predicate.

Given some such $\mathcal{X} = (X, \leq_{\mathcal{X}}, V)$ and $x \in X$ such that $\mathcal{X}, x \vDash_{BS} \varphi$, we choose a set $S \subseteq W$ with the same cardinality as X (an appropriate S can be found since X is finite, while W is countably infinite). Pick any bijection $\iota : W \to X$, and put:

— for any $w, v \in S$: $w \leq_{\mathcal{S}} v$ iff $\iota(w) \leq_{\mathcal{X}} \iota(v)$,
— for any $p \in \Phi$: $[\![p]\!] := \{w \in S \mid \iota(w) \in V(p)\}$.

By definition, $\mathcal{M}$ and $\mathcal{X}$ are isomorphic along the bijection ι. Furthermore, the satisfaction relation $\vDash_{BM}$ defined in Baltag and Smets (2008) runs exactly as the definition of our satisfaction relation $\vDash$.[1] So $\mathcal{X}, x \vDash_{BS} \varphi$ iff $\mathcal{M}, \iota^{-1}(x) \vDash \varphi$. Thus, by the fact that $\mathcal{X}, x \vDash_{BS} \varphi$, it follows that $\mathcal{M}, \iota^{-1}(x) \vDash \varphi$.

This argument shows that every L-consistent sentence $\varphi \in \mathcal{L}$ is satisfiable in a non-empty plausibility model, and from this observation, completeness follows in the usual manner. ⊣

5.2. *Definable Dynamic Attitudes*

5.2.1. Upgrades on Plausibility Models. We have introduced upgrades as transformations of plausibility orders (cf. §1.3). For use in the following, we would like to think of upgrades as applying to plausibility models. Since our setting only models epistemic changes, i.e., changes in the information state of some agent, and not changes of the basic ("ontic", agent-independent) facts of the world, this turns out to be simply a matter of introducing appropriate notation.

Namely, given a plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$, a dynamic attitude $\boldsymbol{\tau}$, and a proposition P, we put

$$\mathcal{M}^{\tau P} := (\mathcal{S}^{\tau P}, [\![\cdot]\!]).$$

So applying the upgrade τP to the plausibility model $\mathcal{M}$ amounts to applying τP to $\mathcal{S}$, and simply "dragging along" the valuation $[\![\cdot]\!]$: the propositions to which atomic sentences evaluate stay the same when applying an upgrade.

5.2.2. Definable Dynamic Attitudes. For the remainder of this chapter, fix two distinct atomic sentences $p_\star, q_\star \in \Phi$, and let $\boldsymbol{\tau}$ be a dynamic attitude.

[1] In other words, our semantics is a notational variant of theirs, with the single difference that the semantics of Baltag and Smets (2008) also allows infinite structures.

A sentence $\vartheta \in \mathcal{L}$ (possibly containing occurrences of $p_\star$ and $q_\star$) *defines* $\boldsymbol{\tau}$ (in $\mathcal{L}$) iff for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ and world $w \in \mathrm{S}$:

$$\mathcal{M}, w \vDash \vartheta \text{ iff } \mathcal{M}^{\boldsymbol{\tau}[\![p_\star]\!]}, w \vDash \Diamond q_\star.$$

The displayed equivalence expresses that ϑ captures precisely the set of worlds satisfying $\Diamond q_\star$ in the plausibility model $\mathcal{M}^{\boldsymbol{\tau}[\![p_\star]\!]}$ resulting from upgrading $\mathcal{M}$ with $\boldsymbol{\tau}[\![p_\star]\!]$. Since we require this to be the case in *all* plausibility models, ϑ can be said to pre-encode the effect of learning $[\![p_\star]\!]$ from a $\boldsymbol{\tau}$-source as far as the defeasible possibility of $[\![q_\star]\!]$ is concerned.

We will show in §5.4 that this is actually tantamount to saying (in a sense to be made precise) that adding operators encoding the dynamic attitude $\boldsymbol{\tau}$ to our language $\mathcal{L}$ does not actually add expressive to $\mathcal{L}$, and this will be crucial to prove completeness for our languages that allow us to talk about dynamic attitudes.

If a sentence ϑ defines a dynamic attitude $\boldsymbol{\tau}$, then we call ϑ a *definition* of $\boldsymbol{\tau}$ (in $\mathcal{L}$). A dynamic attitude $\boldsymbol{\tau}$ is called *definable* (in $\mathcal{L}$) if there exists a definition of $\boldsymbol{\tau}$.

5.2.3. Examples of Definitions. Here are some examples of definitions of dynamic attitudes.

proposition 84.

1. $p_\star \wedge \Diamond(p_\star \wedge q_\star)$ *defines infallible trust* $\mathbf{!}$.
2. $\Diamond(p_\star \wedge q_\star) \vee (\neg p_\star \wedge (\Diamond(\neg p_\star \wedge q_\star) \vee K^{\sim}(p_\star \wedge q_\star)))$ *defines strong trust* $\Uparrow$.
3. $\Diamond q_\star$ *defines neutrality* $\boldsymbol{id}$.
4. $\bot$ *defines isolation* $\varnothing$.

proof.

1. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model, and let $w \in \mathrm{S}$. We have to show that $\mathcal{M}, w \vDash p_\star \wedge \Diamond(p_\star \wedge q_\star)$ iff $\mathcal{M}^{![\![p_\star]\!]}, w \vDash \Diamond q_\star$. Observe that $\mathcal{M}, w \vDash p_\star \wedge \Diamond(p_\star \wedge q_\star)$ iff (by the semantics) $w \in [\![p_\star]\!]$ and there exists $v \in \mathrm{S}$: $v \leq_{\mathcal{S}} w$ and $v \in [\![p_\star]\!] \cap [\![q_\star]\!]$ iff (by definition of !) $w \in \mathrm{S}^{\boldsymbol{\tau}[\![p_\star]\!]}$ and there exists $v \in \mathrm{S}^{\boldsymbol{\tau}[\![p_\star]\!]}$: $v \leq_{\mathcal{S}^{\boldsymbol{\tau}[\![p_\star]\!]}} w$ and $v \in [\![q_\star]\!]$ iff (by the semantics) $\mathcal{M}^{\boldsymbol{\tau}[\![p_\star]\!]}, w \vDash \Diamond q_\star$. This shows our claim, so $p_\star \wedge \Diamond(p_\star \wedge q_\star)$ defines infallible trust $\mathbf{!}$.
2. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model, and let $w \in \mathrm{S}$. We have to show that $\mathcal{M}, w \vDash \Diamond(p_\star \wedge q_\star) \vee (\neg p_\star \wedge (\Diamond(\neg p_\star \wedge q_\star) \vee K^{\sim}(p_\star \wedge q_\star)))$ iff $\mathcal{M}^{\Uparrow[\![p_\star]\!]} \vDash \Diamond q_\star$.

From left to right, suppose that

$$\mathcal{M}, w \vDash \Diamond(p_\star \wedge q_\star) \vee (\neg p \wedge (\Diamond(\neg p_\star \wedge q_\star) \vee K^{\sim}(p_\star \wedge q_\star))).$$

This implies that (1.) or (2.) below holds:

1. $\mathcal{M}, w \vDash \Diamond(p_\star \wedge q_\star)$
2. $\mathcal{M}, w \vDash (\neg p_\star \wedge (\Diamond(\neg p_\star \wedge q_\star) \vee K^{\sim}(p_\star \wedge q_\star)))$.

If (1) holds, then there exists v such that $v \leq_{\mathcal{S}} w$ and $v \in [\![p_\star]\!] \cap [\![q_\star]\!]$. By definition of $\Uparrow$, $v \leq_{\mathcal{S}^{\Uparrow[\![p_\star]\!]}} w$, hence $\mathcal{M}^{\Uparrow[\![p_\star]\!]}, w \vDash \Diamond q_\star$. If (2) holds, then either $\mathcal{M}, w \vDash \neg p_\star \wedge \Diamond \neg p_\star \wedge q_\star$ or (2b) $\mathcal{M}, w \vDash \neg p_\star \wedge K^{\sim}(p_\star \wedge q_\star)$. If (2a) holds, then $\mathcal{M}, w \nvDash p_\star$ and there exists v such that $v \leq_{\mathcal{S}} w$ and $\mathcal{M}, v \nvDash p_\star$ and $\mathcal{M}, v \vDash q_\star$. By definition of $\Uparrow$, $v \leq_{\mathcal{S}^{\Uparrow[\![p_\star]\!]}} w$, hence $\mathcal{M}^{\Uparrow[\![p_\star]\!]}, w \vDash \Diamond q_\star$. If (2b) holds, then $\mathcal{M}, w \nvDash p_\star$ and there exists $v \in S$ such that $\mathcal{M}, v \vDash p_\star \wedge q_\star$. By definition of $\Uparrow$, $v \leq_{\mathcal{S}^{\Uparrow[\![p_\star]\!]}} w$. Hence $\mathcal{M}, w \vDash \Diamond q_\star$. So in either of the two cases (1) or (2), our claim holds, and this concludes the left to right direction.

From right to left, suppose that $\mathcal{M}^{\Uparrow[\![p_\star]\!]}, w \vDash \Diamond q_\star$. Thus there exstis $v \in S$ such that $v \leq_{\mathcal{S}^{\Uparrow[\![p_\star]\!]}} w$ and $v \in [\![q_\star]\!]$. We argue in two cases. First, suppose that $w \in [\![p_\star]\!]$. Then also $v \in [\![p_\star]\!]$ by definition of $\Uparrow$. Furthermore, again by definition of $\Uparrow$, it is the case that $v \leq_{\mathcal{S}} w$. Since $v \in [\![q_\star]\!]$, we conclude that $\mathcal{M}, w \vDash \Diamond(p_\star \wedge q_\star)$. So our claim holds in the first case. Second, suppose that $w \notin [\![p_\star]\!]$. Distinguish two sub-cases. First, supppose that $v \notin [\![p_\star]\!]$. Then $v \leq_{\mathcal{S}} w$ and since $v \in [\![q_\star]\!]$, it follows that $\mathcal{M}, w \vDash \neg p_\star \wedge \Diamond(\neg p_\star \wedge q_\star)$. So our claim holds in the first sub-case. For the second sub-case, suppose that $v \in [\![p_\star]\!]$. Then $v \leq_{\mathcal{S}} w$ and since $v \in [\![q_\star]\!]$, it follows that $\mathcal{M}, w \vDash \neg p_\star \wedge K^{\sim}(p_\star \wedge q_\star)$. So our claim holds in the second sub-case. Thus our claim holds in the second case. In either case, our claim holds, and this concludes the direction from right to left.

3. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model, and let $w \in S$. We have to show that $\mathcal{M}, w \vDash \Diamond q_\star$ iff $\mathcal{M}^{\boldsymbol{id}[\![p_\star]\!]}, w \vDash \Diamond q_\star$. Since $\mathcal{M}^{\boldsymbol{id}[\![p_\star]\!]} = \mathcal{M}$, this is indeed the case.

4. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model, and let $w \in S$. We have to show that $\mathcal{M}, w \vDash \bot$ iff $\mathcal{M}^{\varnothing[\![p_\star]\!]}, w \vDash \Diamond q_\star$. Since neither $\mathcal{M}, w \vDash \bot$ nor $\mathcal{M}^{\varnothing[\![p_\star]\!]}, w \vDash \Diamond q_\star$, the claim holds. ⊣

5.2.4. Non-Definable Attitudes. An example of a dynamic attitude that is, regrettably, *not* definable in $\mathcal{L}$, is minimal trust $\uparrow$. To show this, we introduce an appropriate notion of bisimulation for our epistemic-doxastic language $\mathcal{L}$ (Blackburn, de Rijke, and Venema 2001).

Given a plausibility order $\mathcal{S}$, we write $w \sim_{\mathcal{S}} v$ iff $w, v \in \mathrm{S}$, and we write $w \geq_{\mathcal{S}} v$ iff $v \leq_{\mathcal{S}} w$. Now let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ and $\mathcal{M}' = (\mathcal{S}', [\![\cdot]\!]')$ be plausibility models. A *bisimulation between $\mathcal{M}$ and $\mathcal{M}'$* is a relation $R \subseteq \mathrm{S} \times \mathrm{S}'$ such that for any $w \in \mathrm{S}$, $w' \in \mathrm{S}'$: if wRw', then

— for any $p \in \Phi$: $w \in [\![p]\!]$, iff $w' \in [\![p]\!]'$,
— if there exists v: $w \sim_{\mathcal{S}} v$, then there exists v': $w' \sim_{\mathcal{S}'} v'$ and vRv', and vice versa,
— if there exists v: $w \geq_{\mathcal{S}} v$, then there exists v': $w' \geq_{\mathcal{S}'} v'$ and vRv', and vice versa.

We write $\mathcal{M}, w \simeq \mathcal{M}', w'$ iff there exists a bisimulation R between $\mathcal{M}$ and $\mathcal{M}'$ such that wRw'.

PROPOSITION 85. *Suppose that $\mathcal{M}, w \simeq \mathcal{M}', w'$. Then for any $\varphi \in \mathcal{L}$:*

$$\mathcal{M}, w \vDash \varphi \text{ iff } \mathcal{M}', w' \vDash \varphi.$$

PROOF. Routine (cf. Blackburn et al. (2001)). ⊣

We apply the previous proposition to show that $\uparrow$ is not definable. The proof uses Figure 24: the dotted lines in the Figure indicate a bisimulation between the two plausibility models $\mathcal{M}$ (to the left) and $\mathcal{M}'$ (to the right). Notice that $\mathcal{M}^{\uparrow[\![p]\!]}, z \vDash \Diamond q$, while $(\mathcal{M}')^{\uparrow[\![p]\!]'}, v \nvDash \Diamond q$. However $\mathcal{M}, z \simeq \mathcal{M}', v$, so the two model-world pairs agree, by the above proposition, on *all* sentences in $\mathcal{L}$. Given these facts, a glance at the definition of definability in §5.2.2 is enough to convince us that $\uparrow$ is not definable. We repeat the argument more formally in the next proposition.

PROPOSITION 86. *Minimal trust $\uparrow$ is not definable in $\mathcal{L}$.*

PROOF. Consider the two plausibility models $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ (depicted on the left of Figure 24) and $\mathcal{M}' = (\mathcal{S}', [\![\cdot]\!]')$ (depicted on the right of Figure 24). Clearly, $\mathcal{M}, z \simeq \mathcal{M}', v$. Now let $\vartheta \in \mathcal{L}$ and suppose that $\vartheta \in \mathcal{L}$ defines $\uparrow$ in $\mathcal{L}$. Observe that $\mathcal{M}, z \vDash \vartheta$, since $\mathcal{M}^{\uparrow[\![p_\star]\!]}, z \vDash \Diamond q_\star$. From the fact that $\mathcal{M}, z \simeq \mathcal{M}', v$, it follows by the previous proposition that $\mathcal{M}', v \vDash \vartheta$. Since ϑ defines τ, we conclude that $(\mathcal{M}')^{\tau[\![p_\star]\!]'}, v \vDash \Diamond q_\star$. But the latter statement is false. Hence ϑ does not define $\uparrow$ in $\mathcal{L}$. So $\uparrow$ is not definable in $\mathcal{L}$. ⊣

As a consequence, the setting we consider in this chapter can define only a selection of the important examples of dynamic attitudes considered in this dissertation. The first question this raises is what definability in $\mathcal{L}$ exactly amounts to: just which dynamic attitudes are definable in $\mathcal{L}$? Second, the

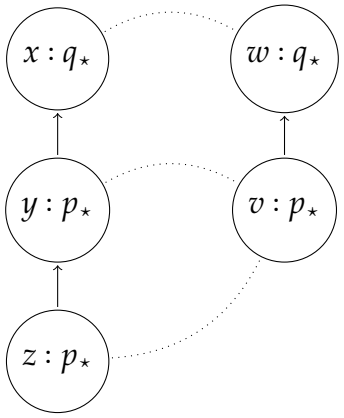


FIGURE 24. Two bisimilar plausibility models: $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ (to the left) and $\mathcal{M}' = (\mathcal{S}', [\![\cdot]\!]')$ (to the right). The labels indicate that $[\![p_\star]\!] = \{y, z\}$, $[\![q_\star]\!] = \{x\}$, $[\![p_\star]\!]' = \{v\}$, $[\![q_\star]\!]' = \{w\}$. The dotted lines indicate a bisimulation between $\mathcal{M}$ and $\mathcal{M}'$.

result points to the fact that the techniques presented in this chapter should ultimately be adapted to *more expressive languages*. One promising candidate is the language which has the modality K, and, in addition, modalities for the strict plausibility order $<_{\mathcal{S}}$, and for the equiplausibility order $\sim_{\mathcal{S}}$ (in a given plausibility order $\mathcal{S}$). As observed in Baltag and Smets (2008), this language is strictly more expressive than the epistemic-doxastic language $\mathcal{L}$ considered here.[2] Third, *weaker* languages might also be considered. For example, suppose we add conditional belief operators $B^{\varphi}\psi$ to propositional logic (interpreted in the natural way, using our definition of the propositional attitude B^P in §1.2.7). The results of van Benthem (2007) essentially show that this language *can* define minimal trust $\uparrow$. The counterexample exposed in Figure 24 does not apply, since the two bisimilar worlds z and v also agree on all sentences in the language with conditional belief operators *after* applying an upgrade $\uparrow P$. Interestingly, then, it is not always the case that a strictly more expressive language can define strictly more dynamic attitudes! The precise formulation and study of a more general notion of definability, applicable to the languages mentioned, is a topic for future research.

[2] In particular, notice that $\Box\varphi$ is equivalent to $[\sim]\varphi \wedge [<]\varphi$.

5.3. *Languages for Definable Sources*

In this section, we construct logical languages for (implicit) agents that receive information from sources that are "definable" in the sense that the dynamic attitude of the agents towards her sources is given by definable dynamic attitudes in the sense of the previous section. We shall assume that each agent has a finite number of such sources at her disposal. Since the dynamic attitude of the agent towards her sources are definable, we can collect suitable definitions for these dynamic attitudes. This leads to the notion of a source structure, which will be an important parameter for our logics.

5.3.1. Definable Sources. A *source structure* is a pair

$$\Sigma = (\mathrm{Att}, \mathrm{Def})$$

such that $\mathrm{Att} \subseteq I$ is a finite, non-empty set of attitude labels,[3] and $\mathrm{Def} : \mathrm{Att} \to \mathcal{L}$ assigns a sentence $\mathrm{Def}(\tau) \in \mathcal{L}$ to each $\tau \in \mathrm{Att}$, in such a way that $\mathrm{Def}(\tau)$ defines τ.

Given a source structure $\Sigma = (\mathrm{Att}, \mathrm{Def})$, a *(definable) source (over* Σ*)* is a pair

$$s := (\tau_s, \vartheta_s),$$

where $\tau_s \in \mathrm{Att}$, and $\vartheta_s = \mathrm{Def}(\tau_s)$. By abuse of notation, we write $s \in \Sigma$ iff s is a source over Σ.

We now use source structures as an additional parameter to construct languages extending our epistemic-doxastic language $\mathcal{L}$.

5.3.2. The extended language $\mathcal{L}[\Sigma]$. For each source structure Σ, we obtain the language $\mathcal{L}[\Sigma]$ by adding a new construction rule to the grammar for $\mathcal{L}$, allowing us to form sentences of the form $[s{:}\,\varphi]\psi$, with $s \in \Sigma$.

Formally, $\mathcal{L}[\Sigma]$ is given by the following grammar:

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid K\varphi \mid [s{:}\,\varphi]\varphi,$$

where $p \in \Phi$, and $s \in \Sigma$.

Read $[s{:}\,\varphi]\psi$ as "after the agent receives the information that φ from a source of type s, ψ holds."

[3] Recall that the index set I is part of the fixed signature we assume as given, cf. §5.1.1.

5.3.3. Semantics of $\mathcal{L}[\Sigma]$. To obtain a semantics for $\mathcal{L}[\Sigma]$, we extend the semantics for $\mathcal{L}$ accordingly. We define, given a plausibility model $\mathcal{M}$:

$$\begin{aligned}
[\![p]\!]_{\mathcal{M}} &:= [\![p]\!] \cap S,\\
[\![\neg\varphi]\!]_{\mathcal{M}} &:= S \setminus [\![\varphi]\!]_{\mathcal{M}},\\
[\![\varphi \wedge \psi]\!]_{\mathcal{M}} &:= [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}},\\
[\![\Box\varphi]\!]_{\mathcal{M}} &:= \Box_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}},\\
[\![K\varphi]\!]_{\mathcal{M}} &:= K_{\mathcal{S}} [\![\varphi]\!]_{\mathcal{M}},\\
[\![[s{:}\,\varphi]\psi]\!] &:= S \cap (S^{\tau_s [\![\varphi]\!]_{\mathcal{M}}} \Rightarrow [\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}}).
\end{aligned}$$

Notice that, except for the last clause, this just repeats the definition of the semantics for $\mathcal{L}$ given in §5.1.4. Notice furthermore that, in the last clause, for any world $w \in S$, we check whether w satisfies the proposition $[\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}}$, provided w is contained in the domain of the upgrade model $\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$. As usual in many dynamic epistemic logics, $[\![[s{:}\,\varphi]\psi]\!]_{\mathcal{M}}$ contains in particular those worlds in S which are *not* contained in $S^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$: these worlds *trivially satisfy* the sentence $[s{:}\,\varphi]\psi$.

As before, we write $\mathcal{M}, w \vDash \varphi$ to mean that $w \in [\![\varphi]\!]_{\mathcal{M}}$. We say that a sentence $\varphi \in \mathcal{L}[\Sigma]$ is *valid* iff $[\![\varphi]\!]_{\mathcal{M}} = S$ for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$. We write $\vDash_{\mathcal{L}[\Sigma]} \varphi$ if φ is valid.

5.3.4. Reintroducing Dynamic Attitudes in the Notation. Given a source structure Σ, a source $s = (\tau_s, \vartheta_s) \in \Sigma$, and sentences $\varphi, \psi \in \mathcal{L}[\Sigma]$, we put

$$[\tau_s \varphi]\psi := [s{:}\,\varphi]\psi.$$

This yields notation familiar from the existing literature (van Benthem 2007, Baltag and Smets 2008), like $[!\varphi]\psi$, $[\Uparrow \varphi]\psi$ (assuming a suitable signature and source structure). But note that, in our setting, $[!\varphi]\psi$ and $[\Uparrow \varphi]\psi$ are not official syntax, but defined notation.

5.4. Expressivity

This section is devoted to showing that for our languages $\mathcal{L}[\Sigma]$, parametrized by a choice of source structure Σ, we can obtain "reduction axioms" in the usual style of dynamic epistemic logic.[4] As a consequence, adding definable sources to $\mathcal{L}$ yields no increase in expressive power to $\mathcal{L}$.

[4] Cf. van Ditmarsch, Kooi, and van der Hoek (2007).

5.4.1. Notation. Let Σ be a source structure, let $s \in \Sigma$, let $\varphi, \psi \in \mathcal{L}[\Sigma]$ be sentences. We write $\varphi[p \mapsto \psi]$ for the sentence resulting from φ by simultaneously substituting all occurrences of p in φ with ψ.

The following notation will be useful:

— $\diamondsuit_s(\varphi, \psi) := \vartheta_s[p_\star \mapsto \varphi, q_\star \mapsto \psi]$,
— $\boxdot_s(\varphi, \psi) := \neg \diamondsuit_s(\varphi, \psi)$,
— $\mathrm{pre}_s(\varphi) := \diamondsuit_s(\varphi, \top)$.

Recalling that $s := (\tau_s, \vartheta_s)$, with τ_s an attitude label, and ϑ_s a definition of $\boldsymbol{\tau_s}$ given by the source structure, we notice that $\diamondsuit_s(\varphi, \psi)$ is the definition of $\boldsymbol{\tau_s}$ (given by Σ) with $p_\star$ substituted by φ and $q_\star$ substituted by ψ.

The notation suggests that our definitions of dynamic attitudes actually serve the purpose of behaving very much like binary modal operators, and this is indeed just what we want to establish in the following.

It is then also natural to define a "dual" $\boxdot_s(\varphi, \psi)$ of $\diamondsuit_s(\varphi, \psi)$, which is just what we do in the second notation introduced above.

Finally, the third piece of notation $\mathrm{pre}_s(\varphi)$ is meant to suggest that using our definitions of dynamic attitudes, we can capture "pre-conditions" (conditions of executability) for applying upgrades, i.e., we would like the sentence $\mathrm{pre}_s(\varphi)$ to capture the worlds (in some given plausibility order $\mathcal{S}$) where the information that φ can be received from the source s.

The remainder of this section is devoted to putting the above notation to work, and making the above informal remarks precise. We start with a simple example.

5.4.2. Example. Using our new notation, we can "rewrite" the usual reduction axioms familiar from dynamic epistemic logic in a generic format. Consider $p_\star \wedge \Diamond(p_\star \wedge q_\star)$, which defines, as we have seen above, the dynamic attitude **!**. Let Σ be a source structure, and let $s \in \Sigma$ be a source such that $\vartheta_s = p_\star \wedge \Diamond(p_\star \wedge q_\star)$ (i.e., ϑ_s defines **!**). Our starting point is the observation that

$$[!p_\star]q_\star \leftrightarrow (p_\star \to q_\star)$$

is a valid sentence of $\mathcal{L}[\Sigma]$. In this equivalence, $p_\star$ on the right hand side functions as the *precondition* of $!p_\star$. We can now "rewrite" the above equivalence using our notation $\mathrm{pre}_s(\varphi)$. Notice that $\mathrm{pre}_s(p_\star)$ works out to $p_\star \wedge \Diamond(p_\star \wedge \top)$, which is equivalent to $p_\star$, true exactly at the worlds contained in $\mathcal{S}^{!\llbracket p_\star \rrbracket}$. So

$$[!p_\star]q_\star \leftrightarrow (\mathrm{pre}_s(p_\star) \to q_\star)$$

is also valid. Now consider a sentence $[!p_\star]\,\Box\, q_\star$. Observe that

$$[!p_\star]\,\Box\, q_\star \leftrightarrow (p_\star \to \Box[!p_\star]q_\star)$$

is valid. This is equivalent to

$$[!p_\star]\,\Box\, q_\star \leftrightarrow (p_\star \to (p_\star \to \Box[!p_\star]q_\star)).$$

As we have seen above, $p_\star$ is equivalent to $\mathrm{pre}_s(p_\star)$. Furthermore, we notice that $p_\star \to \Box[!p_\star]q_\star$ is equivalent to $\boxdot_s(p_\star, [!p_\star]q_\star)$. Substituting equivalents, we obtain that

$$[!p_\star]\,\Box\, q_\star \leftrightarrow (\mathrm{pre}_s(p_\star) \to \boxdot_s(p_\star, [!p_\star]q_\star))$$

is valid. So we can write a reduction law encoding the dynamics of the defeasible knowledge operator $\Box$ under applying updates using the formal machinery introduced so far. While, in the above example, we have only worked with atomic sentences, we now consider laws of this kind more generally.

5.4.3. Reduction Laws. As a matter of general fact, source structures give rise to languages in which "reduction laws" in the usual style of dynamic epistemic logic are valid.

We first observe that, given a source structure Σ, we can write down *preconditions* (conditions of executability) for each source $s \in \Sigma$.

lemma 87. *Let Σ be a source structure, and $s \in \Sigma$. Then for any sentence $\varphi \in \mathcal{L}[\Sigma]$ and plausibility model $\mathcal{M}$: $[\![\mathrm{pre}_s(\varphi)]\!]_{\mathcal{M}} = S^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}$.*

proof. Let Σ be a source structure, and $s \in \Sigma$. Let $\varphi \in \mathcal{L}[\Sigma]$, and let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model. We have to show that $[\![\mathrm{pre}_s(\varphi)]\!]_{\mathcal{M}} = S^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}$. Recall that $\mathrm{pre}_s(\varphi) = \vartheta_s(\varphi, \top)$. Let $[\![\cdot]\!]' := [\![\cdot]\!][p \mapsto [\![\varphi]\!]_{\mathcal{M}}][q \mapsto W]$, and let $\mathcal{N} = (\mathcal{S}, [\![\cdot]\!]')$. By definition of $\mathcal{N}$, we have $[\![\mathrm{pre}_s(\varphi)]\!]_{\mathcal{M}} = [\![\vartheta_s]\!]_{\mathcal{N}}$. Since ϑ_s defines τ_s, we have $[\![\vartheta_s]\!]_{\mathcal{N}} = [\![\Diamond q]\!]_{\mathcal{N}^{\tau_s[\![p]\!]'}}$. By definition of $[\![\cdot]\!]'$: $[\![\vartheta_s]\!]_{\mathcal{N}} = [\![\Diamond \top]\!]_{\mathcal{N}^{\tau_s[\![p]\!]'}}$. By the semantics: $[\![\Diamond \top]\!]_{\mathcal{N}^{\tau_s[\![p]\!]'}} = S^{\tau_s[\![p]\!]'}$. By definition of $[\![\cdot]\!]'$: $S^{\tau_s[\![p]\!]'} = S^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}$. Overall, we have established that $[\![\mathrm{pre}_s(\varphi)]\!]_{\mathcal{M}} = S^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}$, the desired result. ⊣

According to the preceding lemma, given a source $s \in \Sigma$ and a plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$, the sentence $\mathrm{pre}_s(\varphi)$ is satisfied in just those worlds $w \in S$ in which the upgrade $\tau_s[\![\varphi]\!]_{\mathcal{M}}$ is executable (cf. §1.3.2 for the definition of executability). In this sense, $\mathrm{pre}_s(\varphi)$ captures the precondition of $\tau_s[\![\varphi]\!]_{\mathcal{M}}$.

Next, we observe, for use in the proof of Proposition 89 below:

lemma 88. *For any source structure Σ, source $s \in \Sigma$, sentences $\varphi, \psi \in \mathcal{L}[\Sigma]$, plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$, and world $w \in S^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}$:*

$$\mathcal{M}, w \vDash [s{:}\,\varphi]\,\Box\,\psi \quad \textit{iff} \quad \mathcal{M}, w \vDash \boxdot_s(\varphi, [s{:}\,\varphi]\psi)).$$

PROOF. Let Σ be a source structure, $s \in \Sigma$, $\varphi, \psi \in \mathcal{L}[\Sigma]$, $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ a plausibility model, and $w \in S^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$.

We define the plausibility model $\mathcal{N}$ as

$$\mathcal{N} := \mathcal{M}[p \mapsto [\![\varphi]\!]_{\mathcal{M}}, q \mapsto [\![[s{:}\,\varphi]\psi]\!]_{\mathcal{M}}].$$

Now coming from the left,

$$\mathcal{M}, w \vDash [s{:}\,\varphi]\,\Box\,\psi$$

iff (by the semantics and the assumption that $w \in S^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$)

$$\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box\psi$$

iff (by the semantics)

$$\mathcal{S}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box [\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}}.$$

On the other hand, coming from the right,

$$\mathcal{M}, w \vDash \boxdot_s(\varphi, [s{:}\,\varphi]\psi))$$

iff (by definition of $\mathcal{N}$)

$$\mathcal{N}, w \vDash \boxdot_s(p, q)$$

iff (since ϑ_s defines $\boldsymbol{\tau_s}$ and by the assumption that $w \in S^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$)

$$\mathcal{N}^{\tau_s [\![p]\!]_{\mathcal{N}}}, w \vDash \Box q$$

iff (by the definition of $\mathcal{N}$)

$$\mathcal{N}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box q$$

iff (by the semantics)

$$\mathcal{S}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box [\![q]\!]_{\mathcal{N}}.$$

Letting $\mathcal{S}' := \mathcal{S}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}$, to prove our desired equivalence it is thus sufficient to show that $\Box_{\mathcal{S}'} [\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}} = \Box_{\mathcal{S}'} [\![q]\!]_{\mathcal{N}}$. By the semantics, $\Box_{\mathcal{S}'} [\![q]\!]_{\mathcal{N}} = \Box_{\mathcal{S}'}(S' \cap [\![q]\!]_{\mathcal{N}})$. Noticing that, by definition of $\mathcal{N}$, we have that $[\![q]\!]_{\mathcal{N}} = [\![[s{:}\,\varphi]\psi]\!]_{\mathcal{M}}$, it follows using the semantics that $\Box_{\mathcal{S}'} [\![q]\!]_{\mathcal{N}} = \Box_{\mathcal{S}'}(S' \cap (S' \Rightarrow [\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}}))$, hence $\Box_{\mathcal{S}'} [\![q]\!]_{\mathcal{N}} = \Box_{\mathcal{S}'} [\![\psi]\!]_{\mathcal{M}^{\tau_s [\![\varphi]\!]_{\mathcal{M}}}}$, the desired result. ⊣

The preceding lemma is the heart of the proof, in Proposition 89 below, that establishes the validity of the reduction law for defeasible knowledge $\Box$.

PROPOSITION 89. *For any source structure Σ, the following are valid in $\mathcal{L}[\Sigma]$:*

— $[s{:}\,\varphi]p \leftrightarrow (\mathrm{pre}_s(\varphi) \to p)$,

— $[s{:}\,\varphi]\neg\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to \neg[s{:}\,\varphi]\psi)$,
— $[s{:}\,\varphi](\psi \wedge \chi) \leftrightarrow ([s{:}\,\varphi]\psi \wedge [s{:}\,\varphi]\chi)$,
— $[s{:}\,\varphi]K\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to K[s{:}\,\varphi]\psi)$,
— $[s{:}\,\varphi]\,\Box\,\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to \boxdot_s(\varphi, [s{:}\,\varphi]\psi))$.

PROOF. We consider the most interesting item, the last one. Let Σ be a source structure, let $s \in \Sigma$, let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a plausibility model, and let $w \in \mathrm{S}$.

We notice, using Lemma 87, that if $\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}$ is *not* executable in w, then both sides of the bi-implication in (5.) are satisfied at w, i.e., under this assumption, $\mathcal{M}, w \vDash [s{:}\,\varphi]\,\Box\,\psi$ and $\mathcal{M}, w \vDash (\mathrm{pre}_s(\varphi) \to \boxdot_s(\varphi, [s{:}\,\varphi]\psi))$ hold trivially, so our claim holds as well. For the remainder of the proof, we may thus assume that $\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}$ is executable in w. Suppose, then, that $w \in \mathrm{S}^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}$, which, by Lemma 87 is equivalent to saying that $\mathcal{M}, w \vDash \mathrm{pre}_s(\varphi)$. Under this assumption, we observe that

$$\mathcal{M}, w \vDash \mathrm{pre}_s(\varphi) \to \boxdot_s(\varphi, [s{:}\,\varphi]\psi)) \quad \text{iff} \quad \mathcal{M}, w \vDash \boxdot_s(\varphi, [s{:}\,\varphi]\psi)).$$

But the fact that

$$\mathcal{M}, w \vDash [s{:}\,\varphi]\,\Box\,\psi \quad \text{iff} \quad \mathcal{M}, w \vDash \boxdot_s(\varphi, [s{:}\,\varphi]\psi)).$$

is exactly Lemma 88, and we are done. ⊣

5.4.4. COMPLEXITY. Our aim is to show that every sentence $\varphi \in \mathcal{L}[\Sigma]$ is semantically equivalent to a sentence $\psi \in \mathcal{L}$ in the sense that $\vDash_{\mathcal{L}[\Sigma]} \varphi \leftrightarrow \psi$, using the above reduction laws.[5] The proof hinges on an appropriate measure of (syntactic) complexity for sentences of $\mathcal{L}[\Sigma]$, which we proceed to define.

For each source structure Σ, for each sentence $\varphi \in \mathcal{L}[\Sigma]$, the *horizontal depth* $h(\varphi)$ of φ is inductively given by

— $h(p) := 0$
— $h(\neg\varphi) := h(\varphi)$
— $h(\varphi \wedge \psi) := \max\{h(\varphi), h(\psi)\}$
— $h(K\varphi) := h(\varphi)$
— $h(\Box\varphi) := h(\varphi)$
— $h([s{:}\,\varphi]\psi) := 1 + h(\varphi)$

The horizontal depth measures the extent to which dynamic operators are ("horizontally") stacked. For example, $\varphi_1 = [s{:}\,\varphi][s{:}\,\psi]\chi$ is more horizontally complex than $\varphi_2 = [s{:}\,\varphi]\psi$ in the sense that $h(\varphi_1) > h(\varphi_2)$. Intuitively, one

[5] Our strategy for establishing this follows Kooi (2007).

scans a given sentence from left to right, counting dynamic operators in the scope of others.

For each source structure Σ, for each sentence $\varphi \in \mathcal{L}[\Sigma]$, the *vertical depth* $v(\varphi)$ of φ is inductively given by

— $v(p) := 0$
— $v(\neg\varphi) := v(\varphi)$
— $v(\varphi \wedge \psi) := \max\{v(\varphi), v(\psi)\}$
— $v(K\varphi) := v(\varphi)$
— $v(\Box\varphi) := v(\varphi)$
— $v([s{:}\,\varphi]\psi) := \max\{v(\varphi) + 1, v(\psi)\}$

The vertical depth measures the extent to which dynamic operators occur ("vertically") nested inside of other dynamic operators (e.g., $\psi_1 = [s{:}\,[s{:}\,\varphi]\psi]\chi$ is more vertically complex than $\psi_2 = [s{:}\,\varphi]\chi$ in the sense that $v(\psi_1) > v(\psi_2)$). Intuitively, one jumps "inside" the dynamic operators, counting the number of jumps needed until one hits "the bottom of the box."

For each source structure Σ, for each sentence $\varphi \in \mathcal{L}[\Sigma]$, the *complexity* $c(\varphi)$ of φ is given by

$$c(\varphi) := h(\varphi) \cdot v(\varphi).$$

So the complexity of a sentence is obtained by multiplying its horizontal and its vertical depth.

We observe that sentences in $\mathcal{L}$ have complexity 0. Moreover, if a sentence in $\mathcal{L}[\Sigma]$ has complexity 0, then it must be a sentence in $\mathcal{L}$. In a concise statement:

LEMMA 90. *For every source structure Σ, for every sentence $\varphi \in \mathcal{L}[\Sigma]$: $\varphi \in \mathcal{L}$ iff $c(\varphi) = 0$.*

PROOF. A trivial induction on $\varphi \in \mathcal{L}$ shows the left to right direction. For the other direction, suppose that $\varphi \in \mathcal{L}[\Sigma]$, $\varphi \notin \mathcal{L}$. Then φ contains a subformula of the form $[s{:}\,\psi]\chi$. By definition of c, it follows that $c(\varphi) \geq 1$, and this shows the other direction. ⊣

5.4.5. REDUCTION OF $\mathcal{L}[\Sigma]$ TO $\mathcal{L}$. Intuitively, our reduction laws give us the tools to reduce the complexity $c(\varphi)$ of a given sentence φ in $\mathcal{L}[\Sigma]$ until we eventually obtain a sentence ψ of complexity 0, with ψ equivalent to our original φ; by the previous lemma, ψ will be a sentence in $\mathcal{L}$. Let us put this idea into practice.

For $x \in \{v, h, c\}$, we write $\varphi \leq_x \psi$ iff $x(\varphi) \leq x(\psi)$; $\varphi <_x \psi$ iff $x(\varphi) < x(\psi)$; and $\varphi =_x \psi$ iff $x(\varphi) = x(\psi)$.

We first show that, assuming that we *already have* a sentence $\varphi \in \mathcal{L}$, we can always properly reduce the vertical depth of the sentence $[s{:}\psi]\varphi$, for any $\psi \in \mathcal{L}[\Sigma]$, getting rid of the dynamic modality $[s{:}\psi]$, using the reduction laws.

LEMMA 91. *For every source structure Σ, for every sentence $\varphi \in \mathcal{L}$, for every sentence $\psi \in \mathcal{L}[\Sigma]$, there exists a sentence $\chi \in \mathcal{L}[\Sigma]$ such that $\chi <_v [s\psi]\varphi$ and $\vDash_{\mathcal{L}[\Sigma]} \chi \leftrightarrow [s{:}\psi]\varphi$.*

PROOF. Let Σ be a source structure, let $\varphi \in \mathcal{L}$, let $\psi \in \mathcal{L}[\Sigma]$. The proof is by induction on the construction of φ. For each case of the induction, we have to find a sentence $\chi \in \mathcal{L}[\Sigma]$ such that $\chi <_v [s{:}\psi]\varphi$ and $\vDash_{\mathcal{L}[\Sigma]} \chi \leftrightarrow [s{:}\psi]\varphi$.

Consider the case that φ is an atomic sentence. By the reduction law for atomic sentences, $\vDash_{\mathcal{L}[\Sigma]} [s{:}\psi]\varphi \leftrightarrow \mathrm{pre}_s(\varphi) \to \psi$, and clearly, $\mathrm{pre}_s(\varphi) \to \psi <_v [s{:}\psi]\varphi$, so we have found the desired χ.

We now assume, as our induction hypothesis, that we have shown the claim for all subformulas of φ.

The four cases of the inductive step are all similar, so we restrict ourselves to discussing two cases: negation and defeasible knowledge.

Consider the case that φ is of the form $\neg\rho$. By the reduction law for negation, $\vDash_{\mathcal{L}[\Sigma]} [s{:}\psi]\neg\rho \leftrightarrow (\mathrm{pre}_s(\psi) \to \neg[s{:}\psi]\rho)$. Since ρ is a subformula of $\neg\rho$, there exists a sentence ϑ such that $\vDash_{\mathcal{L}[\Sigma]} \vartheta \leftrightarrow [s{:}\psi]\rho$ and $\vartheta <_v [s{:}\psi]\rho$. Clearly, $(\mathrm{pre}_s(\psi) \to \neg\vartheta) <_v [s{:}\psi]\neg\rho$. But observe that $\vDash_{\mathcal{L}[\Sigma]} (\mathrm{pre}_s(\psi) \to \neg\vartheta) \leftrightarrow [s{:}\psi]\neg\rho$. Hence we have found the desired χ.

Consider now the case of defeasible knowledge. Suppose that φ is of the form $\Box\rho$. By the reduction law for defeasible knowledge, $\vDash_{\mathcal{L}[\Sigma]} [s\psi]\Box\rho \leftrightarrow (\mathrm{pre}_s(\psi) \to \boxdot_s(\psi, [s\psi]\rho))$. Since ρ is a subformula of $\Box\rho$, by the induction hypothesis, there exists a sentence β such that $\beta <_v [s\psi]\rho$ and $\vDash_{\mathcal{L}[\Sigma]} \beta \leftrightarrow [s\psi]\rho$. Clearly, $(\mathrm{pre}_s(\psi) \to \boxdot_s(\psi, \beta)) <_v [s\psi]\Box\rho$. But observe that $\vDash_{\mathcal{L}[\Sigma]} (\mathrm{pre}_s(\psi) \to \boxdot_s(\psi, \beta)) \leftrightarrow [s\psi]\Box\rho$. Hence we have found the desired χ. ⊣

Using the previous lemma, we show that any sentence in $\mathcal{L}[\Sigma]$ can be reduced to a sentence in $\mathcal{L}$.

PROPOSITION 92 ("Semantic Reduction Theorem"). *For every source structure Σ, for every sentence $\varphi \in \mathcal{L}[\Sigma]$ there exists a sentence $\varphi^\# \in \mathcal{L}$ such that $\vDash_{\mathcal{L}[\Sigma]} \varphi \leftrightarrow \varphi^\#$.*

PROOF. The proof is by induction on $c(\varphi)$. If $c(\varphi) = 0$, then $\varphi \in \mathcal{L}$ (cf. Lemma 90), and we are done. Assume now that $c(\varphi) > 0$ and suppose that we have shown the claim for all sentences $\varphi' \in \mathcal{L}[\Sigma]$ such that $\varphi' <_c \varphi$. Choose a subformula ψ of φ with $h(\psi) = 1$ (some such ψ exists, for otherwise, $c(\varphi) = 0$, contradiction). The sentence ψ is a Boolean combination of $\mathcal{L}$-sentences and sentences of the form $[s{:}\alpha]\vartheta$. Choose some such $[s{:}\alpha]\vartheta$ (again, some such

$[s{:}\alpha]\vartheta$ exists, for otherwise, $h(\psi) = 0$, contradiction). Since $h(\psi) = 1$, it follows that $h([s{:}\alpha]\vartheta) = 1$, so $h(\vartheta) = 0$, i.e., $\vartheta \in \mathcal{L}$. By the previous lemma, there exists a sentence χ equivalent to $[s{:}\alpha]\vartheta$ such that $\chi <_v [s{:}\alpha]\vartheta$. Thus $\chi <_c [s{:}\alpha]\vartheta$. Since $[s{:}\alpha]\vartheta \leq_c \varphi$ ($[s{:}\alpha]\vartheta$ is a subformula of φ!), it follows that $\chi <_c \varphi$. By the induction hypothesis, χ is equivalent to an $\mathcal{L}$-sentence. So $[s{:}\alpha]\vartheta$ is equivalent to an $\mathcal{L}$-sentence. Since $[s{:}\alpha]\vartheta$ was arbitrarily chosen, it follows that ψ is equivalent to an $\mathcal{L}$-sentence. Since ψ was also arbitrarily chosen, this establishes that for all subformulas ψ of φ such that $h(\psi) = 1$, there exists a sentence ψ' such that $h(\psi') = 0$ and $\vDash_{\mathcal{L}[\Sigma]} \psi \leftrightarrow \psi'$. Substituting each ψ' for the corresponding ψ in our original φ yields a formula φ' that is equivalent to φ. Moreover, $\varphi' <_h \varphi$ (indeed, $h(\varphi') = h(\varphi) - 1$), and thus $\varphi' <_c \varphi$. By the induction hypothesis, φ' is equivalent to an $\mathcal{L}$-sentence, say $\varphi^{\#}$. But as we have observed above, $\vDash_{\mathcal{L}[\Sigma]} \varphi \leftrightarrow \varphi'$. So φ is equivalent to the $\mathcal{L}$-sentence $\varphi^{\#}$, which finishes the inductive step of the proof, and we are done. ⊣

As a corollary, we obtain that adding definable sources to $\mathcal{L}$ does not add expressive power.

5.4.6. Expressivity. Given two logical languages $\mathcal{L}_1$ and $\mathcal{L}_2$ interpreted in plausibility models, we say that $\mathcal{L}_1$ is *at least as expressive as* $\mathcal{L}_2$ iff for every sentence $\varphi \in \mathcal{L}_1$, there exists a sentence $\psi \in \mathcal{L}_2$ such that $[\![\varphi]\!]_{\mathcal{M}} = [\![\psi]\!]_{\mathcal{M}}$ for any plausibility model $\mathcal{M}$. We say that $\mathcal{L}_1$ and $\mathcal{L}_2$ are *co-expressive* if $\mathcal{L}_1$ is at least as expressive as $\mathcal{L}_2$, and $\mathcal{L}_2$ is at least as expressive as $\mathcal{L}_1$.

theorem 93. *For every source structure Σ: $\mathcal{L}$ and $\mathcal{L}[\Sigma]$ are co-expressive.*

proof. The fact that $\mathcal{L}[\Sigma]$ is at least as expressive as $\mathcal{L}$ follows from the fact that $\mathcal{L}[\Sigma]$ extends $\mathcal{L}$, and the semantics of $\mathcal{L}[\Sigma]$ agrees with the semantics of $\mathcal{L}$ for sentences in $\mathcal{L}$. The fact that $\mathcal{L}$ is at least as expressive as $\mathcal{L}[\Sigma]$ is a corollary of Proposition 92. ⊣

We turn to a converse of sorts to the previous theorem. Namely, we show that extending the language $\mathcal{L}$ with a *non-definable* dynamic attitude increases the expressive power. While this is intuitively obvious, working it out in slightly more detail serves to make the case that the notion of definability we have introduced really *characterizes* co-expressivity, as per the conjunction of Theorem 93 above, and Proposition 94 below.

Let τ be a dynamic attitude which is not definable (for short: a non-definable dynamic attitude). The language $\mathcal{L}[\tau]$ is obtained by adding a formation rule to $\mathcal{L}$ allowing us to build sentences of the form $[\tau\varphi]\psi$. The

semantics for $\mathcal{L}[\tau]$ extends the semantics for $\mathcal{L}$, i.e., we add the following clause:

$$[\![[\tau\varphi]\psi]\!]_{\mathcal{M}} = [\![\psi]\!]_{\mathcal{M}^{\tau[\![\varphi]\!]_{\mathcal{M}}}}.$$

Now we observe:

PROPOSITION 94. *$\mathcal{L}[\tau]$ is more expressive than $\mathcal{L}$.*

PROOF. Since τ is not definable, there exists no sentence $\vartheta \in \mathcal{L}$ such that for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ and world $w \in \mathrm{S}$: $\mathcal{M}, w \vDash \vartheta$ iff $\mathcal{M}^{\tau[\![p]\!]_{\mathcal{M}}}, w \vDash \Diamond q$. However, $\mathcal{M}^{\tau[\![p]\!]_{\mathcal{M}}}, w \vDash \Diamond q$ iff $\mathcal{M}, w \vDash \langle\tau p\rangle q$. So there exists no sentence $\vartheta \in \mathcal{L}$ such that for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ and world $w \in \mathrm{S}$: $\mathcal{M}, w \vDash \vartheta$ iff $\mathcal{M}, w \vDash \langle\tau p\rangle q$. Hence $\mathcal{L}[\tau]$ is more expressive than $\mathcal{L}$. ⊣

So adding definable attitudes to $\mathcal{L}$ yields no increase in expressive power, while adding undefinable attitudes to $\mathcal{L}$ does yield an increase in expressive power. In this sense, the notions of definability and expressivity match.

5.5. Completeness

This section supplies an axiomatization of the language $\mathcal{L}[\Sigma]$ interpreted over plausibility models.

5.5.1. THE LOGIC OF DEFINABLE SOURCES. For every source structure Σ, *the logic* $\mathrm{L}[\Sigma]$ *of definable sources* is given by adding the axioms below to the logic of defeasible and indefeasible knowledge L (the additional axioms are just the "reduction laws" discussed above, copied from the statement of Proposition 89 for easy reference):

— $[s{:}\,\varphi]p \leftrightarrow (\mathrm{pre}_s(\varphi) \to p)$,
— $[s{:}\,\varphi]\neg\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to \neg[s{:}\,\varphi]\psi)$,
— $[s{:}\,\varphi](\psi \wedge \chi) \leftrightarrow ([s{:}\,\varphi]\psi \wedge [s{:}\,\varphi]\chi)$,
— $[s{:}\,\varphi]K\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to K[s{:}\,\varphi]\psi)$,
— $[s{:}\,\varphi]\,\Box\,\psi \leftrightarrow (\mathrm{pre}_s(\varphi) \to \boxdot_s(\varphi, [s{:}\,\varphi]\psi))$.

We now show that $\mathrm{L}[\Sigma]$ is weakly sound and complete.

PROPOSITION 95 ("Syntactic Reduction Theorem"). *For every source structure Σ, for every sentence $\varphi \in \mathcal{L}[\Sigma]$: there exists a sentence $\psi \in \mathcal{L}$ such that φ and ψ are provably equivalent in* $\mathrm{L}[\Sigma]$.

PROOF. The proof is analogous to the proof of the "semantic reduction theorem" (Proposition 92): first, show the syntactic analogue of Lemma 91 using the reduction axioms. Then, prove the statement of the syntactic reduction theorem by induction on $c(\varphi)$, using the fact that our logic allows substitution of equivalents.[6] ⊣

THEOREM 96 (Completeness). *For every source structure* Σ*:* $\mathrm{L}[\Sigma]$ *is weakly sound and complete w.r.t. plausibility models.*

PROOF. For soundness, it suffices to show that our axioms are valid, and that the rules of inference preserve validity. As observed earlier, the axioms of the logic of defeasible and indefeasible knowledge L are indeed valid, and the rules of inference do indeed preserve validity. Furthermore, we have shown in Proposition 89 that the reduction axioms are valid. Soundness follows by induction on the length of a derivation in $\mathrm{L}[\Sigma]$. For completeness, we argue as follows: suppose that $\varphi \in \mathcal{L}[\Sigma]$ is valid. By Proposition 95, there exists a sentence $\varphi^t \in \mathcal{L}$ such that $\varphi \leftrightarrow \varphi^t$ is provable in $\mathrm{L}[\Sigma]$. By soundness of $\mathrm{L}[\Sigma]$, $\varphi \leftrightarrow \varphi^t$ is valid, so $\varphi^t \in \mathcal{L}$ is valid. By completeness of L, φ^t is provable in L. Since $\mathrm{L}[\Sigma]$ extends L, φ^t is also provable in $\mathrm{L}[\Sigma]$. Also, we know from above that $\varphi^t \to \varphi$ is provable in $\mathrm{L}[\Sigma]$. So φ is provable in L (using modus ponens). Hence any valid sentence $\varphi \in \mathcal{L}[\Sigma]$ is provable in $\mathrm{L}[\Sigma]$. So $\mathrm{L}[\Sigma]$ is complete. ⊣

5.6. Logics for Mutual Trust and Distrust

In this section, we generalize our results to the multi-agent setting discussed in §2.7 and §2.8 of this dissertation. The presentation parallels the one given for the single-agent case in §§5.1–5.5, so we proceed at a faster pace.

5.6.1. MULTI-AGENT PLAUSIBILITY MODELS. A *(multi-agent) plausibility model* is a pair

$$\mathcal{M} = (\{\mathcal{S}_a\}_{a \in \mathcal{A}}, [\![\cdot]\!])$$

where $\{\mathcal{S}_a\}_{a \in \mathcal{A}}$ is a multi-agent plausibility order, and $[\![\cdot]\!]$ is a valuation.

[6] Cf. Kooi (2007). The point is that if a sentence φ is provable in $\mathrm{L}[\Sigma]$, and φ contains an occurrence of some sentence ψ as a subformula, and, moreover, ψ and some sentence ψ' are provably equivalent in $\mathrm{L}[\Sigma]$, then we may replace ψ with ψ' in φ to obtain another sentence φ' that is provable in $\mathrm{L}[\Sigma]$.

5.6.2. The Epistemic-Doxastic Language $\mathcal{L}[\mathcal{A}]$. The language $\mathcal{L}[\mathcal{A}]$ (called the *(multi-agent) epistemic-doxastic language*) is given by the following grammar ($p \in \Phi$):

$$\varphi \quad ::= \quad p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid \Box_a\varphi,$$

where $p \in \Phi$, and $a \in \mathcal{A}$ ($a \neq b$). Read $K_a\varphi$ as *agent a infallibly (*or: *indefeasibly) knows that φ*; read $\Box_a\varphi$ as *agent a defeasibly knows that φ*.

5.6.3. Semantics. We interpret the language $\mathcal{L}[\mathcal{A}]$ in the way familiar from the preceding section. Let $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$ be a multi-agent plausibility model. Then we define by recursion on sentence structure:

$$\begin{array}{lcl}
[\![p]\!]_{\mathcal{M}} & := & [\![p]\!] \cap \mathrm{S}, \\
[\![\neg\varphi]\!]_{\mathcal{M}} & := & \mathrm{S} \setminus [\![\varphi]\!]_{\mathcal{M}}, \\
[\![\varphi \wedge \psi]\!]_{\mathcal{M}} & := & [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}}, \\
[\![\Box_a\varphi]\!]_{\mathcal{M}} & := & \{v \in \mathrm{S} \mid v \in \Box_{\mathcal{S}_{a(v)}} [\![\varphi]\!]_{\mathcal{M}}\}, \\
[\![K_a\varphi]\!]_{\mathcal{M}} & := & \{v \in \mathrm{S} \mid v \in K_{\mathcal{S}_{a(v)}} [\![\varphi]\!]_{\mathcal{M}}\}.
\end{array}$$

As earlier, we write $\mathcal{M}, w \vDash \varphi$ to mean that $w \in [\![\varphi]\!]_{\mathcal{M}}$. We say that a sentence $\varphi \in \mathcal{L}[\mathcal{A}]$ is *valid* iff $[\![\varphi]\!]_{\mathcal{M}} = \mathrm{S}$ for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$. And we write $\vDash_{\mathcal{L}[\mathcal{A}]} \varphi$ if φ is valid.

5.6.4. The Multi-Agent Logic of Defeasible and Indefeasible Knowledge. The multi-agent logic of defeasible and indefeasible knowledge is just the obvious analogue of the single-agent logic of defeasible and indefeasible knowledge considered in the previous chapter. Formally, $\mathrm{L}[\mathcal{A}]$ is given by the following axioms and rules:

— *Axioms:*
  — All instances of theorems of propositional calculus
  — $K_a(\varphi \to \psi) \to (K_a\varphi \to K_a\psi)$
  — $\Box_a(\varphi \to \psi) \to (\Box_a\varphi \to \Box_a\psi)$
  — The S5 axioms for K_a
  — The S4 axioms for $\Box_a$
  — $K_a\varphi \to \Box_a\varphi$
  — $K_a(\varphi \vee \Box_a\psi) \wedge K_a(\psi \vee \Box_a\varphi) \to (K_a\varphi \vee K_a\psi)$
— *Rules of inference:*
  — From φ and $\varphi \to \psi$ infer ψ
  — From φ infer $K_a\varphi$ and $\Box_a\varphi$

The axiomatization of $\mathrm{L}[\mathcal{A}]$ is again due to Baltag and Smets (2008).

THEOREM 97 (Baltag and Smets (2008)). *The multi-agent logic of defeasible and indefeasible knowledge* $\mathrm{L}[\mathcal{A}]$ *is weakly sound and complete w.r.t. multi-agent plausibility models.*

PROOF. Analogous to the proof of Theorem 83. ⊣

5.6.5. THE LANGUAGE $\mathcal{L}[\mathcal{A}, \Sigma]$. For each source structure Σ, we obtain the language $\mathcal{L}[\mathcal{A}, \Sigma]$ by adding two new construction rules to the grammar for $\mathcal{L}[\mathcal{A}]$, allowing us to build sentences of the form $s_{a\to b}$, with $s \in \Sigma$ a source, and $a, b \in \mathcal{A}$, and sentences of the form $[a{:}\,\varphi]\psi$, with $a \in \mathcal{A}$ an agent. As detailed below, the additional syntactic material will allow us to study communication acts made by agents.

Formally, the language $\mathcal{L}[\mathcal{A}, \Sigma]$ is given by the following grammar:

$$\varphi \quad ::= \quad p \mid s_{a\to b} \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid K_a\varphi \mid \Box_a\varphi \mid [a{:}\,\varphi]$$

where $p \in \Phi$, $a, b \in \mathcal{A}$ $(a \neq b)$, and $s \in \Sigma$.

Read $[a{:}\,\varphi]\psi$ as "after the communication act $a{:}\,\varphi$, ψ holds", and read $s_{a\to b}$ as "agent a considers agent b to be a source of type s". By the latter reading, we mean that the attitude of agent a towards agent b is given by the attitude label τ_s corresponding to the dynamic attitude $\boldsymbol{\tau_s}$.

Right away, we notice that there is a "syntactic mismatch" in that a source $s := (\tau_s, \vartheta_s)$ over Σ supplies us with a sentence $\tau_s \in \mathcal{L}$, while $\mathcal{L}[\mathcal{A}, \Sigma]$ extends $\mathcal{L}[\mathcal{A}]$ (the multi-agent version) rather than $\mathcal{L}$ (the single-agent version). This issue is easily resolved; we will take care of it in §5.6.8 below.

5.6.6. TRUST-PLAUSIBILITY MODELS OVER SOURCE STRUCTURES. In what kind of structure would we like to interpret the language $\mathcal{L}[\mathcal{A}, \Sigma]$? The following notion is obviously useful towards answering the question.

A *trust-plausibility model* is a triple $(\{\mathcal{S}\}_{a\in\mathcal{A}}, T, [\![\cdot]\!])$ such that $(\{\mathcal{S}\}_{a\in\mathcal{A}}, T)$ is a multi-agent trust-plausibility order, and $(\{\mathcal{S}\}_{a\in\mathcal{A}}, [\![\cdot]\!])$ is a multi-agent plausibility model.

However, we are here interested in a specific kind of trust-plausibility model. Namely, we are interested in trust-plausibility models where all the mutual dynamic attitudes of the agents come from a given source structure Σ, in the sense that if, according to the trust labeling, at some world w, some agent a has the dynamic attitude σ to some agent b, then there better exist some $s \in \Sigma$ such that $\boldsymbol{\tau_s} = \sigma$. This ensures that all the attitudes agents entertain towards each other are actually definable.

So let Σ be a source structure.

— A trust graph T is a *trust graph over* Σ iff for each $a, b \in \mathcal{A}$ such that $a \neq b$ there exists a source $s \in \Sigma$ such that $T(a,b) = \boldsymbol{\tau_s}$.
— Let $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ be a multi-agent plausibility order. A trust labeling T over $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ is a *trust labeling over* Σ *and* $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$ iff for each $w \in \mathrm{S}$: T_w is a trust graph over Σ.
— A trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T, [\![\cdot]\!])$ is a *trust-plausibility model over* Σ iff T is a trust labeling over Σ and $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$.

A trust-plausibility model over Σ is thus a trust-plausibility model such that for any world w and for any pair of agents $a, b \in \mathcal{A}$ such that $a \neq b$: the attitude $T_w(a,b)$ "corresponds" to a source $s \in \Sigma$ in the sense that $T_w(a,b) = \boldsymbol{\tau_s}$.

5.6.7. Semantics. The language $\mathcal{L}[\mathcal{A},\Sigma]$ is interpreted in trust-plausibility models over Σ.

As a preliminary: given a trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, [\![\cdot]\!], T)$, and a communication act $[a{:}\,P]$, we write $\mathcal{M}^{[a:P]}$ for the trust-plausibility model

$$(\{\mathcal{S}_a[a{:}\,P]_{a\in\mathcal{A}}, [\![\cdot]\!], T).$$

So applying a communication act $[a{:}\,P]$ to $\mathcal{M}$ amounts to applying $[a{:}\,P]$ to the underlying trust-plausibility order $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$, and dragging the trust labeling T and the valuation $[\![\cdot]\!]$ given by $\mathcal{M}$ along, leaving both unchanged (cf. §2.7.8 for the definition of $\mathcal{S}_a[a{:}\,P]_{a\in\mathcal{A}}$).

Now we define, by recursion on sentence structure, for each source structure Σ, for each trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T, [\![\cdot]\!])$ over Σ and sentence $\varphi \in \mathcal{L}[\mathcal{A},\Sigma]$:

$$\begin{array}{lcl}
[\![p]\!]_{\mathcal{M}} & := & [\![p]\!] \cap \mathrm{S}, \\
[\![s_{a\to b}]\!]_{\mathcal{M}} & := & \{w \in \mathrm{S} \mid T_w(a,b) = \boldsymbol{\tau_s}\}, \\
[\![\neg\varphi]\!]_{\mathcal{M}} & := & \mathrm{S} \setminus [\![\varphi]\!]_{\mathcal{M}}, \\
[\![\varphi \wedge \psi]\!]_{\mathcal{M}} & := & [\![\varphi]\!]_{\mathcal{M}} \cap [\![\psi]\!]_{\mathcal{M}}, \\
[\![\Box_a\varphi]\!]_{\mathcal{M}} & := & \{v \in \mathrm{S} \mid v \in \Box_{\mathcal{S}_{a(v)}} [\![\varphi]\!]_{\mathcal{M}}\}, \\
[\![K_a\varphi]\!]_{\mathcal{M}} & := & \{v \in \mathrm{S} \mid v \in K_{\mathcal{S}_{a(v)}} [\![\varphi]\!]_{\mathcal{M}}\}, \\
[\![[a{:}\,\varphi]\psi]\!]_{\mathcal{M}} & := & \mathrm{S} \cap (\mathrm{S}[a{:}\,[\![\varphi]\!]_{\mathcal{M}}] \Rightarrow [\![\psi]\!]_{\mathcal{M}[a:[\![\varphi]\!]_{\mathcal{M}}]}).
\end{array}$$

As before, we write $\mathcal{M}, w \vDash \varphi$ to mean that $w \in [\![\varphi]\!]_{\mathcal{M}}$. We say that a sentence $\varphi \in \mathcal{L}[\mathcal{A}]$ is *valid* iff $[\![\varphi]\!]_{\mathcal{M}} = \mathrm{S}$ for any plausibility model $\mathcal{M} = (\mathcal{S}, [\![\cdot]\!])$. And we write $\vDash_{\mathcal{L}[\mathcal{A},\Sigma]} \varphi$ if φ is valid.

5.6.8. Notation. For every agent a, we introduce the following recursive syntactic translation $\cdot^a$ from sentences of $\mathcal{L}$ to sentences of $\mathcal{L}[\mathcal{A}]$:

$$p^a := p, \quad (\neg\varphi)^a := \neg(\varphi^a), \quad (\varphi \wedge \psi)^a := \varphi^a \wedge \psi^a, \quad (K\varphi)^a := K_a\varphi^a, \quad (\Box\varphi)^a := \Box_a\varphi^a.$$

As one can see by inspecting the clauses of the definition, all that $\cdot^a$ does is label all occurrences of the symbols "K" and "$\Box$" in a given sentence with a, i.e., replacing K with K_a, and $\Box$ with $\Box_a$. We observe:

lemma 98. *For any $\varphi \in \mathcal{L}$, for any agent $a \in \mathcal{A}$, for any trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T, [\![\cdot]\!])$, for any $w \in \mathrm{S}$:*

$$\mathcal{M}, w \vDash \varphi^a \textit{ iff } (\mathcal{S}_{a(w)}, [\![\cdot]\!]), w \vDash \varphi$$

proof. Easy induction on $\varphi \in \mathcal{L}$. ⊣

Next, we define abbreviations paralleling the ones we have introduced in the previous chapter (cf. §5.4.1) to the present multi-agent setting. Given a source structure Σ, and a definition $s = (\tau_s, \vartheta_s)$, recall that ϑ_s is an $\mathcal{L}$-sentence that defines $\boldsymbol{\tau_s}$. Putting $\vartheta_s^a := (\vartheta_s)^a$, we "personalize" ϑ_s for each agent $a \in \mathcal{A}$.

Now we make the following abbreviations:

— $\diamond_b(a\colon\varphi, \psi) := \bigwedge_{s\in\Sigma}(s_{b\to a} \to \vartheta_s^b(\varphi, \psi))$,
— $\boxdot_b(a\colon\varphi, \psi) := \neg \diamond_b (a\colon\varphi, \neg\psi)$,
— $\mathrm{pre}(a\colon\varphi) := \bigwedge_{b\neq a}(\diamond_b(a\colon\varphi, \top))$.

Again (cf. the remarks in §5.4.1), $\diamond_b(a\colon\varphi, \psi)$ intentionally looks very much like a binary modality, with $\boxdot_b(a\colon\varphi, \psi)$ its "dual." The aim is to capture, in some current trust-plausibility model, what an agent b defeasibly knows after the communication act $[a\colon\varphi]$ is applied to that model. The sentence $\mathrm{pre}(a\colon\varphi)$, on the other hand, is meant to capture the precondition (condition of executability) of the communication act $[a\colon\varphi]$. We verify that these abbreviations fulfill their purpose in the next paragraph.

5.6.9. Reduction Laws. Establishing the desired reduction laws can now be done analogously to §5.4.3, where we considered the reduction laws for the single-agent case. Since the overall setup is somewhat different, we do provide details.

lemma 99. *For any source structure Σ, for any sentence $\varphi \in \mathcal{L}[\mathcal{A}, \Sigma]$, for any trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_b\}_{b\in\mathcal{A}}, T, [\![\cdot]\!])$ over Σ, for any $w \in \mathrm{S}$, for any $a \in \mathcal{A}$:*

$$[\![\mathrm{pre}(a\colon\varphi)]\!]_{\mathcal{M}} = \mathrm{S}[a\colon\varphi].$$

PROOF. Let Σ be a source structure, let $\varphi \in \mathcal{L}[\mathcal{A}, \Sigma]$, let $\mathcal{M} = (\{\mathcal{S}_b\}_{b\in\mathcal{A}}, T, [\![\cdot]\!])$ be a trust-plausibility model over Σ, let $w \in \mathrm{S}$, let $a \in \mathcal{A}$. Consider now the following chain of equivalences:

$$\mathcal{M}, w \vDash \mathrm{pre}(a{:}\,\varphi)$$

iff (unfolding the abbreviation $\mathrm{pre}(a{:}\,\varphi)$)

$$\mathcal{M}, w \vDash \bigwedge_{b\neq a} \bigwedge_{s\in\Sigma} s_{b\to a} \to \vartheta^b_s(\varphi, \top)$$

iff (by the semantics and the definition of $\vartheta^b_s(\varphi, \top)$)

$$\forall b \neq a \in \mathcal{A} \forall s \in \Sigma : \text{ if } T_w(b,a) = \boldsymbol{\tau_s}, \text{ then } \mathcal{M}[p \mapsto [\![\varphi]\!]_{\mathcal{M}}, q \mapsto W], w \vDash \vartheta^b_s$$

iff (by Lemma 98)

$$\forall b \neq a \in \mathcal{A} \forall s \in \Sigma : \text{ if } T_w(b,a) = \boldsymbol{\tau_s}, \text{ then } (\mathcal{S}_{b(w)}, [\![\cdot]\!][p \mapsto [\![\varphi]\!]_{\mathcal{M}}, q \mapsto W]), w \vDash \vartheta_s$$

iff (by the semantics)

$$\forall b \neq a \in \mathcal{A} \forall s \in \Sigma : \text{ if } T_w(b,a) = \boldsymbol{\tau_s}, \text{ then } (\mathcal{S}_{b(w)}, [\![\cdot]\!][p \mapsto [\![\varphi]\!]_{\mathcal{M}}]), w \vDash \vartheta_s(p, \top)$$

iff (by definition of $\mathrm{pre}_s(p)$)

$$\forall b \neq a \in \mathcal{A} \forall s \in \Sigma : \text{ if } T_w(b,a) = \boldsymbol{\tau_s}, \text{ then } (\mathcal{S}_{b(w)}, [\![\cdot]\!][p \mapsto [\![\varphi]\!]_{\mathcal{M}}]), w \vDash \mathrm{pre}_s(p)$$

iff (by Lemma 87 and the definition of $[\![\cdot]\!][p \mapsto [\![\varphi]\!]_{\mathcal{M}}]$)

$$\forall b \neq a \in \mathcal{A} \forall s \in \Sigma : \text{ if } T_w(b,a) = \boldsymbol{\tau_s}, \text{ then } w \in (\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}$$

iff (by the fact that $\mathcal{M}$ is a trust-plausibility model over Σ)

$$\forall b \neq a \in \mathcal{A} : w \in (\mathcal{S}_{b(w)})^{\boldsymbol{\tau}^w_{b\to a}[\![\varphi]\!]_{\mathcal{M}}}$$

iff (by definition of a communication act)

$$w \in \mathrm{S}[a{:}\,\varphi].$$

This proves our claim. ⊣

LEMMA 100. *For any source structure Σ, sentences $\varphi, \psi \in \mathcal{L}[\mathcal{A}, \Sigma]$, for any trust-plausibility model $\mathcal{M} = (\{\mathcal{S}_b\}_{b\in\mathcal{A}}, T, [\![\cdot]\!])$, world $w \in \mathrm{S}[a{:}\,\varphi]$ and agents $a, b \in \mathcal{A}$:*

$$\mathcal{M}, w \vDash [a{:}\,\varphi] \Box_b \psi \quad \textit{iff} \quad \mathcal{M}, w \vDash \boxdot_b(a{:}\,\varphi, [a{:}\,\varphi]\psi).$$

PROOF. Let Σ be a source structure, let $\varphi, \psi \in \mathcal{L}[\Sigma, \mathcal{A}]$, let $\mathcal{M} = (\{\mathcal{S}_b\}_{b\in\mathcal{A}}, T, [\![\cdot]\!])$ be a trust-plausibility model, let $w \in \mathrm{S}[a{:}\,\varphi]$, let $a, b \in \mathcal{A}$.

We define the trust-plausibility model $\mathcal{N}$ as

$$\mathcal{N} := \mathcal{M}[p \mapsto [\![\varphi]\!]_{\mathcal{M}}, q \mapsto [\![[a{:}\,\varphi]\psi]\!]_{\mathcal{M}}].$$

We use (cf. §2.7.7) the notation $\mathcal{S}[a{:}\,\varphi]$ for the natural product order on $\mathrm{S}[a{:}\,\varphi]$, that is $\mathcal{S}[a{:}\,\varphi] := (\mathrm{S}[a{:}\,\varphi], \mathrm{S}[a{:}\,\varphi] \times \mathrm{S}[a{:}\,\varphi])$. To simplify the notation, suppose that $s \in \Sigma$ is such that $T_w(b, a) = \boldsymbol{\tau_s}$.

Coming from the left,

$$\mathcal{M}, w \vDash [a{:}\,\varphi]\,\Box_b\,\psi$$

iff (by the semantics and the assumption that $w \in \mathrm{S}[a{:}\,\varphi]$)

$$\mathcal{M}[a{:}\,\varphi], w \vDash \Box_b\psi$$

iff (by the semantics and the definition of a communication act)

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}} \cap \mathcal{S}[a{:}\,\varphi], w \vDash \Box[\![\psi]\!]_{\mathcal{M}[a:\varphi]}$$

iff (using the definition of $\mathcal{S}[a{:}\,\varphi]$)

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box(\mathrm{S}[a{:}\,\varphi] \Rightarrow [\![\psi]\!]_{\mathcal{M}[a:\varphi]}).$$

On the other hand, coming from the right,

$$\mathcal{M}, w \vDash \boxdot_b(\varphi, [a{:}\,\varphi]\psi))$$

iff (by definition of $\mathcal{N}$)

$$\mathcal{N}, w \vDash \boxdot_b(p, q)$$

iff (by definition of $\boxdot_b(p, q)$ and Lemma 98, using the fact that $T_w(b, a) = \boldsymbol{\tau_s}$)

$$\mathcal{S}_{b(w)}, w \vDash \boxdot_s(p, q)$$

iff (since ϑ_s defines $\boldsymbol{\tau_s}$ and $w \in \mathrm{S}[a\varphi]$)

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![p]\!]_{\mathcal{N}}}, w \vDash \Box[\![q]\!]_{\mathcal{N}}$$

iff (by definition of $\mathcal{N}$)

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box[\![q]\!]_{\mathcal{N}}.$$

It is thus sufficient to show that

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box(\mathrm{S}[a{:}\,\varphi] \Rightarrow [\![\psi]\!]_{\mathcal{M}[a:\varphi]}) \quad \text{iff} \quad (\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box[\![q]\!]_{\mathcal{N}}.$$

To prove this, we argue as follows:

$$(\mathcal{S}_{b(w)})^{\boldsymbol{\tau_s}[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box(\mathrm{S}[a{:}\,\varphi] \Rightarrow [\![\psi]\!]_{\mathcal{M}[a:\varphi]})$$

iff (by the semantics)

$$(\mathcal{S}_{b(w)})^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box(\mathrm{S}[a{:}\,\varphi] \Rightarrow (\mathrm{S}[a{:}\,\varphi] \cap [\![[a{:}\,\varphi]\psi]\!]_{\mathcal{M}}))$$

iff (by propositional reasoning)

$$(\mathcal{S}_{b(w)})^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box(\mathrm{S}[a{:}\,\varphi] \Rightarrow [\![[a{:}\,\varphi]\psi]\!]_{\mathcal{M}})$$

iff (using Lemma 99)

$$(\mathcal{S}_{b(w)})^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box[\![\mathrm{pre}(a{:}\,\varphi) \to [a{:}\,\varphi]\psi]\!]_{\mathcal{M}}$$

iff (by the semantics)

$$(\mathcal{S}_{b(w)})^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box[\![[a{:}\,\varphi]\psi]\!]_{\mathcal{M}}$$

iff (by definition of $\mathcal{N}$)

$$(\mathcal{S}_{b(w)})^{\tau_s[\![\varphi]\!]_{\mathcal{M}}}, w \vDash \Box[\![q]\!]_{\mathcal{N}}.$$

This is the desired result. ⊣

PROPOSITION 101. *For any source structure Σ: the following are valid in $\mathcal{L}[\mathcal{A}, \Sigma]$:*

— $[a{:}\,\varphi]p \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to p)$
— $[a{:}\,\varphi]s_{b\to c} \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to s_{b\to c})$
— $[a{:}\,\varphi]\neg\varphi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to \neg[a{:}\,\varphi]\varphi)$
— $[a{:}\,\varphi](\psi \wedge \chi) \leftrightarrow ([a{:}\,\varphi]\psi \wedge [a{:}\,\varphi]\chi)$
— $[a{:}\,\varphi]K_b\psi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to K_b[a{:}\,\varphi]\psi)$
— $[a{:}\,\varphi]\,\Box_b\,\psi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to \boxdot_b(a{:}\,\varphi, [a{:}\,\varphi]\psi))$

PROOF. We consider the most interesting item, the last one. Let Σ be a source structure, let $\mathcal{M} = (\{\mathcal{S}_a\}_{a\in\mathcal{A}}, T, [\![\cdot]\!])$ be a trust-plausibility model over Σ, and let $w \in \mathrm{S}$. We notice, using Lemma 99, that if $w \notin \mathrm{S}[a{:}\,\varphi]$, then both sides of the bi-implication are satisfied at w, i.e., under this assumption, $\mathcal{M}, w \vDash [a{:}\,\varphi]\,\Box_b\,\psi$ and $\mathcal{M}, w \vDash (\mathrm{pre}(a{:}\,\varphi) \to \boxdot_b(a{:}\,\varphi, [a{:}\,\varphi]\psi))$ hold trivially, so our claim holds as well. We may thus assume that $w \in \mathrm{S}[a{:}\,\varphi]$, which, by Lemma 99 is equivalent to saying that $\mathcal{M}, w \vDash \mathrm{pre}(a{:}\,\varphi)$. So we merely need to show that, under this assumption, $\mathcal{M}, w \vDash [a{:}\,\varphi]\,\Box_b\,\psi$ iff $\mathcal{M}, w \vDash \boxdot_b(a{:}\,\varphi, [a{:}\,\varphi]\psi)$, but this is exactly Lemma 100. ⊣

5.6.10. Logics of Mutual Trust and Distrust. For every source structure Σ, *the logic of mutual trust and distrust* $\mathrm{L}[\Sigma, \mathcal{A}]$ is obtained by adding the axioms below to the multi-agent logic of defeasible and indefeasible knowledge $\mathrm{L}[\mathcal{A}]$ (the recursion axioms are just copied from the statement of Proposition 101):

— *Recursion Axioms:*

- $[a{:}\,\varphi]p \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to p)$
- $[a{:}\,\varphi]s_{b\to c} \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to s_{b\to c})$
- $[a{:}\,\varphi]\neg\varphi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to \neg[a{:}\,\varphi]\varphi)$
- $[a{:}\,\varphi](\psi \wedge \chi) \leftrightarrow ([a{:}\,\varphi]\psi \wedge [a{:}\,\varphi]\chi)$
- $[a{:}\,\varphi]K_b\psi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to K_b[a{:}\,\varphi]\psi)$
- $[a{:}\,\varphi]\,\Box_b\,\psi \leftrightarrow (\mathrm{pre}(a{:}\,\varphi) \to \boxdot_b(a{:}\,\varphi, [a{:}\,\varphi]\psi))$

— *Trust Axioms:*

- $\bigwedge_{a\in\mathcal{A}} \bigwedge_{b\neq a} (\bigvee_{s\in\Sigma} s_{a\to b})$
- $\bigwedge_{a\in\mathcal{A}} \bigwedge_{b\neq a} \bigwedge_{s\in\Sigma} \bigwedge_{s'\neq s} (s_{a\to b} \to \neg s'_{a\to b})$

The two trust axioms capture our assumption that every agent has a unique attitude towards each other agent (cf. §2.7.5 for more discussion).

We establish the completeness of the logic of mutual trust and distrust by adapting earlier results. We provide a sketch.

theorem 102. *For any source structure Σ: the logic of trust and distrust $\mathrm{L}[\Sigma, \mathcal{A}]$ is weakly sound and complete w.r.t. trust-plausibility models over Σ.*

proof. *(Sketch.)* The soundness half works as in earlier results (cf., e.g., Theorem 83). For completeness, let Σ be a source structure. Our reduction axioms allow us to reduce the language $\mathcal{L}[\mathcal{A}, \Sigma]$ to the language built over the grammar

$$\varphi \quad ::= \quad p \mid s_{a\to b} \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid \Box_a\varphi,$$

with $p \in \Phi$, $a \in \mathcal{A}$, $s \in \Sigma$. We refer to this language as $\mathcal{L}[\mathcal{A}+]$. We refer to the proof system obtained by dropping the reduction axioms from $\mathrm{L}[\mathcal{A}, \Sigma]$ as $\mathrm{L}[\mathcal{A}+]$. The fact that every sentence in $\mathcal{L}[\mathcal{A}, \Sigma]$ is provably equivalent to a sentence in $\mathcal{L}[\mathcal{A}+]$ can be shown by adapting the proof of Theorem 95, using the reduction axioms. Call this adapted result the "syntactic reduction theorem for $\mathcal{L}[\mathcal{A}, \Sigma]$". The proof that $\mathrm{L}[\mathcal{A}+]$ is weakly sound and complete w.r.t. trust-plausibility models over Σ is obtained by adapting the argument in the proof of Theorem 2.5 in Baltag and Smets (2008). Finally, we deduce, as in the proof of Theorem 96, the completeness of $\mathrm{L}[\mathcal{A}+]$ from the completeness of $\mathrm{L}[\mathcal{A}, \Sigma]$, using the syntactic reduction theorem for $\mathrm{L}[\mathcal{A}, \Sigma]$. ⊣

Chapter 6.
Epistemic Modals

The work of this chapter begins with a change of perspective, as it introduces a second interpretation of the notion of a dynamic attitude.

So far in this dissertation, we have understood dynamic attitudes as formal correlates of reliability assessments—capturing in what way, and to what extent, a speaker trusts or distrusts a source. Reliability assessments are, conceptually speaking, part of the epistemic state of an agent. The overall idea we have pursued was that the agent uses her reliability assessment towards a particular source to determine how to change her epistemic state when receiving information from that source. The aspects of the epistemic state of an agent that we have been concerned with were thus, in the single-agent case, (1) her beliefs about the world, modeled by a (single-agent) plausibility order, and (2) her attitudes towards sources, modeled by dynamic attitudes.[1] The informational inputs that trigger belief changes, on the other hand, were understood as representable by means of their propositional content, as sets of possible worlds.

In this chapter, we shall view dynamic attitudes not as part of the epistemic state of an agent, but as *part of the informational input received by the agent*. From this changed perspective, the input received by the agent is not simply a proposition, but a proposition embedded under an operator. Our proposal is that an upgrade τP given by applying a dynamic attitude τ to a proposition P can be used to give a semantics for epistemic modal sentences of the form $\textsc{modal}(p)$, as in "it might be the case that p".

Let me motivate the perspective change further with an example. Recall first the set of scenarios discussed in the introduction of this dissertation, repeated below. We have considered various ways in which I may receive the information that there are tigers in the Amazon jungle:

[1] In the multi-agent setting we have studied in §§2.7–2.8 and §5.6, the single-agent plausibility order lives inside a "bigger" structure, a multi-agent plausibility order, also representing the agent's information about *other* agents.

(1) a. I read a somewhat sensationalist coverage in the yellow press claiming this.

b. I read a serious article in a serious newspaper claiming this.

c. I read the Brazilian government officially announcing that tigers have been discovered in the Amazon area.

d. I see a documentary on TV claiming to show tigers in the Amazon jungle.

e. I read an article in Nature by a famous zoologist reporting of tigers there.

f. I travel to the Amazon jungle, and see the tigers.

Examples of this kind have been used to motivate the concept of a dynamic attitude (cf. the introduction). The information that there are tigers in the Amazon jungle may be received from a variety of sources; and since we trust these sources to varying degrees, our "epistemic response", i.e., the way we change our epistemic state on receiving the information that there are tigers in the Amazon jungle, will differ depending on the particular source.

Let us now consider a different type of scenario. Suppose that I talk to a trusted friend about wildlife in the Amazon jungle. Consider six variants of what my friend might tell me:

(2) a. There might be tigers in the Amazon jungle.

b. There could be tigers in the Amazon jungle.

c. There may be tigers in the Amazon jungle.

d. There should be tigers in the Amazon jungle.

e. There must be tigers in the Amazon jungle.

f. There are tigers in the Amazon jungle.

In all six cases, I receive information about tigers in the Amazon jungle. Also, in all six cases, I receive information about tigers in the Amazon jungle from the same source, my trusted friend. But in all six cases, the way I change my epistemic state on receiving that information from that source seems to be different. Since the source of information is the same in each case, the fact that the information change is different would seem to be due to the fact that *the information received* is different: the epistemic modal auxiliary (might, could, should, etc) modulates the "information uptake" on the side of the recipient.

As a result, depending on which modal is used, I will adopt different stances towards the proposition that there are tigers in the Amazon jungle.

For a preliminary diagnosis of the flavour of these different epistemic stances, a first observation to make is that the assertions lower in the list tend to make the assertions higher in the list redundant. We can see this by composing pairs of the statements: in (a) below, the second assertion is informative after the first; but not so in (b):

(3) a. There may be tigers in the Amazon jungle. In fact, there *are* tigers in the Amazon jungle!

b. There are tigers in the Amazon jungle. ?In fact, there *may be* tigers in the Amazon jungle![2]

Also, consider this variation on the data in (2), pointing in the same direction:

(4) a. Scientists found out that there may be tigers in the Amazon jungle. Then, they found out that there are tigers in the Amazon jungle.

b. Scientists found out that there are tigers in the Amazon jungle. ?Then, they found out that there may be tigers in the Amazon jungle.

The sequence in (a) reports progresss; the sequence in (b) is hard to make sense of.

Another observation is that I may cite my friend's assertion that there *are* tigers in the Amazon jungle to justify my belief that there are tigers there. But it would be highly odd to cite his assertion that there *may* be tigers in the Amazon jungle in the same way:

(5) a. I believe that there are tigers in the Amazon jungle, because my friend told me that there are tigers in the Amazon jungle.

b. ?I believe that there are tigers in the Amazon jungle, because my friend told me that there may be tigers in the Amazon jungle.

The assertion that there are tigers in the Amazon jungle is thus stronger (in some relevant, not yet precise sense) than the assertion that there may be tigers in the Amazon jungle; and this is reflected in the change in my epistemic state induced by integrating each sentence into my epistemic state.

[2] As usual, the judgement diacritic "?" indicates contextual infelicity of the sentence labeled with it.

So epistemic modality in natural language and assessments of reliability as we have studied them so far share an important property: a notion of strength seems to apply to them in a similar way.

Another upshot of the discussion so far is that, in natural language use, information change is not simply induced by transmitting propositions from source to recipient. Rather, we may distinguish between different ways in which the same proposition may be embedded under a modal operator. Intuitively, the modal operator seems to modulate the "force" with which the epistemic state of the recipient is changed.

Phrasing this in terms of our formal setting, the picture is then that epistemic states of single agents (recipients of information) may still be captured by (single-agent) plausibility orders; however, the role assumed by dynamic attitudes changes. Given a plausibility order $\mathcal{S}$, an upgrade τP will take $\mathcal{S}$ to some new order $\mathcal{S}^{\tau P}$. The suggestion is now, as indicated above, that τ may figure as the semantic correlate of some epistemic modal. So dynamic attitudes can be seen as supplying a semantics for epistemic modals.

This picture will be immediately recognizable for anyone familiar with the dynamic semantics literature. There, the idea is that "knowing the meaning of a sentence is knowing the change it brings about in the cognitive state of anyone who wants to incorporate the information conveyed by it." (Veltman 2005) Here, we study a specific type of sentence, which we assume to have the logical form

$$\textsc{modal}(p),$$

with MODAL some modal auxiliary, and p a sentence embedded under the former. The proposal here is that the semantics of such a sentence can be given in terms of an upgrade

$$\tau P,$$

where τ is a dynamic attitude capturing the meaning of MODAL, and P is a proposition capturing the meaning of p.

The following sections develop our account in more detail. We focus here on the modal auxiliaries which have wide-spread uses with an epistemic reading. This excludes epistemic uses of *would* that seem to occur only in the context of the so-called "epistemic would equatives" (Ward, Birner, and Kaplan 2003):

(6) a. Who is the man with the microphone?

b. That would be the Dean.

It also excludes the modal auxiliary *can* which is generally assumed to only admit non-epistemic readings (Hacquard 2011), expressing, for example, an ability:

(7) a. John can do the job.

b. Mary can lift 200 pounds.

We are then left with *must*, *may*, *might*, *should* and *could*. We shall, as a rule, ignore other readings of these modals (for example, deontic readings of *should*), as they fall outside of the scope of the framework developed in this dissertation.

Besides the modals themselves, the question which semantics should be assigned to non-modalized indicatives is pertinent, based on the hypothesis that epistemic modal statement live on an ascending scale of certainty which is *topped* by categorical, flat-out statements. In other words (as we shall argue below): modalized claims induce a weaker form of acceptance in the hearer than a categorical claim that something *is* the case. While in its general outlines, this assumption is not a matter of debate in the literature, particular details are controversial, and we will discuss those below.

Our aim is thus to provide a semantics for sentences of the form

$$\textsc{accept}(p)$$

in terms of upgrades τP, where τ is a dynamic attitude understood as interpreting

$$\textsc{accept} \in \{\text{is}, \text{must}, \text{should}, \text{may}, \text{could}, \text{might}\},$$

and P is a proposition interpreting what is traditionally called the "prejacent", the sentence p embedded under the modal.

We shall generally assume that a "typical" sentence p is satisfied in all and only the P-worlds (same character, upper case), where the worlds in P are drawn from some arbitrary but fixed set of possible worlds W.

The following analysis tries to accumulate empirical and formal arguments, so that in the end, operations can be identified that plausibly capture the meaning of the members of ACCEPT. In the end, we do not actually arrive at one proposal, but at a selection of alternative ones. This underlines the exploratory nature of this chapter, whose main point is to make the case that the setting developed in this dissertation may provide an interesting toolbox for formal semantics. We will approach the issue from three directions. §6.1 contrasts our approach with two main approaches in the literature, the proposal that epistemic modals express *degrees of commitment* on behalf of the

speaker, and the proposal that epistemic modals are contextually restricted quantifiers. §6.2 adopts a "bottom-up" perspective, identifying distinctions between the various "acts of acceptance" that epistemic modals correspond to on our account; and §6.3 proceeds in a more "top-down" fashion, by considering intuitions about what *all* acts of acceptance considered here have in common.

6.1. Epistemic Modals as Dynamic Operators

A key intuition voiced above as well as in the literature is that different modal expressions carry a distinct *modal force*. On the view developed here, epistemic modals correspond to types of transformations of plausibility orders, with the latter representing the epistemic state of an agent processing a modalized sentence. The central claim is the following: epistemic modals transform the hearer's plausibility assessment of the prejacent. The force of a modal is then characterized by the extent to which accepting a modal sentence $\textsc{modal}(p)$ transforms the recipient's information state. It should be already clear at this point that the framework developed in the preceding chapters lends itself well to arriving at a formal conception of modal force. In fact, our setting offers two perspectives on the notion. First, the force of a modal τ can be understood in terms of the fixed point $\overline{\tau}$ of τ (cf. §1.7): a modal σ is "more forceful" than another modal τ if the fixed point of σ *entails* the fixed point of τ, which is to say that whenever $\overline{\sigma}P$ is satisfied, then so is $\overline{\tau}P$. But we can also understand modal force directly in terms of the transformations themselves: σ is "more forceful" than τ iff σ *subsumes* τ (cf. §1.8). This means that after applying an upgrade σP, applying an upgrade τP is redundant (for example: after accepting that *John must be in London*, hearing that *John might be in London* provides no new information). By Theorem 11 both perspectives amount to the same thing.

The idea of cashing in the notion of modal force in terms of one transformation making another redundant seems very plausible (in the introduction to this chapter, we have already used this idea implicitly). The idea is also inherent in previous work on modality in dynamic frameworks, even though previous research has generally tended to (1) solely focus on the epistemic modals *must* and *might* (Veltman 1996, Willer 2013, forthcoming), and (2) study modals as dynamic operators, without making the connection to propositional attitudes explicit.

Before developing the view in more formal detail, this section argues that the position defended here has empirical bite, and allows us to make sense of

certain empirical phenomena that otherwise remain obscure. I will do so by contrasting the proposal with two important perspectives on epistemic modality in the literature: the view (popular in the descriptive linguistics tradition) according to which epistemic modals express degrees of commitment, and the view (predominant in formal semantics) according to which epistemic modals are contextually restricted quantifiers.

6.1.1. Degrees of Commitment. According to a wide-spread view, "epistemic modality in natural language marks the degree and/or source of the speaker's commitment to the embedded proposition."[3]

This proposal suggests, for example, that a speaker's communicating that there may be tigers in the Amazon jungle expresses a *weaker* commitment on behalf of the speaker than a speaker's saying, flat-out, that there are tigers in the Amazon jungle. Epistemic modals are thus assigned a clear job description: they are to be used to communicate degrees of (epistemic) commitment. This proposal is very attractive in that it provides us with an immediate explanation of modals having particular forces lying on a scale: the force of a modal simply resides in the degree of commitment of the speaker that the modal expresses. Consequently, the strength relations between modals derive from the fact that expressing a particular degree of strength is, essentially, what modals are supposed to do.

The view has problems, however (cf. Papafragou (2006) for more discussion). The following issue is the central one for our purposes. Consider the following example. Bob is watching the second season of *Homeland*. Alice has already seen all episodes, but is keeping Bob company. At some point, the following dialogue unfolds:

(8) a. *Bob:* Could Galvez be a mole?

b. *Alice:* He could be.

Let us assume that the commitment view on epistemic modals is correct. Then the first sentence uttered by Alice reveals, it seems safe to say, some rather weak form of commitment, on Alice's behalf, to the fact that Galvez is a mole. One gets the impression that she regards it as a remote possibility that Galvez is a mole, at the very least: a possibility not to be excluded.

Actually, however, this is only part of what Alice says, as I have omitted the last bit of the dialogue. In fact, the dialogue runs as follows:

[3]This quote is taken from Papafragou (2006), who, however, does not endorse this view, but merely reports it, and is actually critical of it. As cited by Papafragou, the view is endorsed, for example, by Halliday (1970), Palmer (1990) and Bybee and Fleischman (1995).

(9) a. *Bob:* Could Galvez be a mole?

b. *Alice:* He could be. But as a matter of fact, he isn't.

Continuing to assume that the commitment view is correct, we are now in a bit of trouble. Let's see why. Clearly, the second sentence uttered by Alice reveals that she is commited to Galvez *not* being a mole. Combining this with the previous observation that Alice regards it as a remote plausibility that Galvez is a mole, Alice is contradicting herself. However, I find no reason to take issue with what Alice says: it does not seem to me that she is contradicting herself. And this indicates that the commitment view is wrong.

Notice also that the coherence of Alice's order is sensitive to the order. The following reply to Bob's question is marked (i.e., stands out as unusual and uncommon, "feels wrong") and sounds rather incoherent:

(10) *Alice:* He isn't. ??But as a matter of fact he could be.

The following variants do not sound better:

(11) a. *Alice:* He isn't. ?But he could be.

b. *Alice:* As a matter of fact he isn't. ?But he could be.

The speaker commitment approach provides no clue why this should be so; we will return to this issue below.

6.1.2. Modals as Restricted Quantifiers. Another perspective on epistemic modals is provided by the predominant view in theoretical linguistics originating in Kratzer's work (Kratzer 1981, 2012). On this view, epistemic modals carry truth-conditional content on the one hand, but are context-dependent on the other hand.[4] The context, essentially, provides an ordering on a set of possible worlds, which in turn is derived from a number of propositions (a "premise set"). For an example, let such an order $\leq$ be given (usually, $\leq$ is assumed to be a reflexive and transitive but not necessarily total relation: some worlds may be incomparable). The main idea is then that a sentence like *there may be tigers in the Amazon jungle* is true iff, among the *best* worlds in $\leq$ (the worlds w such that we can find no v with $v < w$), some worlds are worlds in which there are tigers in the Amazon jungle. So *may* functions

[4] In fact, Kratzer's account provides a general framework for dealing not only with epistemic, but with any kind of modality in natural language. But for present purposes, it is solely the epistemic readings of modals that matter, given our limited scope in this chapter.

as a "restricted quantifier", i.e., an existential quantifier quantifying over the best worlds in the order.

The account assumes that in linguistic practice, context supplies a set of premises from which the order $\leq$ can be "projected". For example, in a certain context, it may be the current state of scientific knowledge about the Amazon jungle that figures as a conversational background, and the above sentence would then be paraphrasable as:

(12) In view of the current state of scientific knowledge, there may be tigers in the Amazon jungle.

Since the latter paraphrase suggests that the speaker herself endorses the view that there may be tigers in the Amazon jungle, it has usually been assumed that, at least in the majority of cases, the evidence of the speaker should be considered part of the conversational background (the "speaker inclusion constraint"), and, in fact, the so-called "speaker-centric" reading, in which it is *solely* the speaker's evidence that matters, is sometimes taken to be the default reading of an epistemic modal sentence.

The Kratzerian account offers a simple account of modal force. Since modal sentences carry truth-conditions, it is simply truth-conditional entailment that is responsible for modal force: that *must* has a stronger force than *may* is due to the fact that whenever it is true that something must be the case it is also true that that same thing may be the case.

Returning to the above discussion, the first observation to be made is that the Kratzerian account can make sense of the data that turned out to be puzzling for the speaker commitment approach above. Namely, it could be argued from this perspective that Alice's statement (that Galvez could be a mole, but as a matter of fact isn't a mole) is contextually decoded in the following way:

(13) In view of your evidence, he could be a mole. But in view of *my* evidence, he isn't.

This seems to express in a fairly intuitive way what the discourse in (9b) intuitively communicates. While Bob's evidence does not yet exclude that Galvez is a mole, Alice is in a more privileged epistemic position. Remember that she has already seen all episodes. And since, in later episodes, it turns out that Galvez is not a mole, he is not a mole in view of Alice's evidence. This rather straightforward account favours the more flexible Kratzerian approach over the speaker commitment view discussed earlier.

Notice, however, that the quantificational view does not explain why reversing the order would matter for the markedness of the discourse:

(14) In view of my evidence, he isn't a mole. But in view of your evidence, he could be a mole.

(14) is not more marked than (13), which is unexpected, given that, as observed above, (10) and (9b) differ strongly w.r.t. markedness. Since the whole idea was that (14) elicits the contextual meaning of (10), and (13) the contextual meaning of (9b), this result is unsatisfactory.

Notice, on the other hand, that there is nothing problematic about the initial example from the perspective of the dynamic account sketched so far. A speaker can easily accomodate the information that Galvez is not a mole after first accepting that she could be. This happens all the time: whenever we gain information, the space of options shrinks and previously possible, or plausible options come to be impossible, or implausible. The dynamic account also explains why reversing the order should make a difference: having discarded the possibility that Galvez is a mole (after accepting Alice's first statement), the hearer has no room anymore to accomodate the possibility that he *could be* a mole. So from the perspective of any hearer accepting both statements in turn, the discourse must be incoherent.[5]

Furthermore, on the dynamic account, a simple pragmatic explanation for the intuition that epistemic modals reflect speaker commitments is possible. In a context where an information exchange may be presumed by the participants to proceed in a cooperative fashion in the sense of Grice, it is natural to assume that the information change induced in the hearer by accepting an epistemic modal sentence is matched by a corresponding propositional attitude of the speaker.[6] In other words, if I come to believe that John must be in London based on information obtained from you, it is usually safe to assume that you, as the source of information, also believe that John must be in London.

The context considered in the *Homeland* example, however, is not a standard case of a cooperative information exchange: being as informative as possible is clearly not guaranteed to be the most helpful thing to do when watching a TV series. Rather, it is customarily regarded as spoiling the fun. In this sense, the "adversarial" strategy of revealing as little information as

[5]This type of consideration was a main point of Veltman (1996)'s work on *might*, and is not a novelty of the current proposal.

[6]In fact, this assumption is closely related to the notion of *honesty* discussed in an earlier chapter of this dissertation, cf. §2.8.

possible may be the *really* cooperative one in such instances. In such circumstances, speakers need not be committed themselves to the epistemic modal claims they are making.[7]

Overall, we may conclude that in a standard, Gricean context, speakers will indeed tend to express degrees of commitment by epistemic modal talk, and they will aim at converting the hearer to the *same* degree of commitment—as predicted by the commitment view. But the dynamic account also handles deviant cases with ease.

The view advocated here has another advantage over the Kratzerian account, which I discuss next. Noticeably, on Kratzer's view, epistemic modals *report about* what is compatible with, plausible from the perspective of, or entailed by the information of a group. It seems to me that this leads to conceptual problems. Consider the following example.

(15) a. *Peggy:* Don must be in his office.

b. *Peggy:* I believe that Don is in his office.

Let us assume that the context supplies a speaker-centric reading for Peggy's utterance that Don must be in his office. The rough truth-conditions for the sentence (15a) on a Kratzerian analysis would then run as follows: the sentence is true at a world w iff it follows from Peggy's evidence in w that Don is in his office. Now what about (15b), where Peggy is reporting her belief that Don is in his office? It seems that the truth conditions of the sentence *I believe that Don is in his office* uttered by Peggy in the same context amount to the very same thing. It seems, then that the two sentences are equivalent (in this context). However, the following is only felicitous as a reply to (15b), not to (15a):

(16) Yes, I know. But *I* don't believe it. I have seen him in the conference room.

This observation seems hard to make sense of from the point of view of the Kratzerian framework. The dynamic account, on the other hand, does not face a problem. Here is a sketch of an analysis: Peggy's saying that Don must be in his office amounts, when accepted, to increasing the plausibility of worlds where Don is in his office over the plausibility of worlds where he isn't (just in what particular way is less important for the concrete example: details follow below). On the other hand, Peggy's saying that she believes that Don is

[7] Cf. Verbrugge and Mol (2008) for a more systematic perspective on the relationship between cooperative and adversarial communication.

in his office communicates, when accepted, that worlds in which Peggy does *not* believe that Don is in his office may be excluded from consideration. And, of course, one may agree that worlds where Peggy believes that Don is not in his office may be excluded from consideration without finding it plausible that Don is in his office.

The advantages of the dynamic account that have been highlighted so far are thus, in my view, the following: (1) it allows us to preserve the plausible intuition that epistemic modals are usually related to the speaker's beliefs in one way or another; (2) it avoids the pitfall posed by the observation that using a modal does not necessarily require a particular degree of commitment (the *Homeland* example); (3) it allows us to make sense of the order-dependence of modal discourse (as has already been observed by Veltman); (4) it allows us to draw a clear distinction between modals and reports about the propositional attitude of a speaker (the *Peggy* example).

A final point to make is that our account, with its emphasis on *fixed points* of dynamic transformations, can be seen as a quite conservative "remodeling" of Kratzer's approach. In the Kratzerian tradition, epistemic modals essentially correspond to what has been called propositional attitudes in this thesis. The claim underlying my proposal—very much in the overall spirit of this dissertation—is that rather than *corresponding* to propositional attitudes, epistemic modals *realize* propositional attitudes in the hearer, as the dynamic output of accepting an epistemic modal claim.

6.2. *The Modal Scale*

A main source of evidence about epistemic modals are speaker judgements about entailment relations among modalized sentences. On balance, these judgments suggest that the group of modals we are concerned with here *live on a scale* that is topped by the force of a flat-out assertion that something *is* the case. Let us look into the matter in more detail.

6.2.1. *Is* vs. *Must*. Consider the following pair of sentences:

(17) a. John must be in London.

b. John is in London.

The conviction that (17a) has a more "tentative" flavour than (17b) has been voiced many times in the literature. Recently, von Fintel and Gillies

have challenged this position.[8] Von Fintel and Gillies claim that saying that something must be the case is just as strong as saying that something is the case, the difference being that *must* carries an evidential signal to the extent that the claim is based on an indirect inference. And, indeed, *must*-claims cannot be based on direct inference:

(18) *Looking at the rain pouring down:* ??It must be raining.

However, the arguments put forward by von Fintel and Gillies leave open the possibility that English tends to present claims that are marked as derived from indirect evidence as "weak".[9] The question is what the character of this weakness should amount to. I find it quite difficult to judge whether there can be situations where it is *true* that a particular state of affairs *must* obtain, while it is false that this particular state of affairs *does* obtain (or vice versa, for that matter). That judgements of this kind are difficult may point to the fact that we are not trained to think about epistemic modals in terms of their truth conditions.[10]

The position suggested by the account presented here is that the difference between *must* and *is* lies in the fact that claims of the former kind are marked as defeasible, while claims of the latter kind are not. Consider the following example:

(19) a. John must be in his office. Let's go see if he's there.

b. John is in his office. ?Let's go see if he's there.

If John is in his office, there is no need to check whether he is here. Suggesting such a check seems pointless; so 19b sounds strange. But 19a does not sound strange at all. In particular, it does not sound like the speaker is retreating to a weaker position, as he would be by saying:

(20) John must be in his office. If you don't believe me, let's go see if he's there.

In other words: if one is convinced that something must be the case, one may still find oneself in a position where one wants to check whether that

[8] Cf. von Fintel and Gillies (2010), who extensively reference the literature in support of the "mantra", as they call it.

[9] Just as a judge in a court would tend to regard an elaborate chain of inferences in support of the fact that John murdered Bill as weak compared to an eyewitness account of the actual incident of John strangling Bill.

[10] Cf. Yalcin (2011), Willer (2013) for elaborations on this point.

something indeed *is* the case. And this seems to me to point to the fact that the "tentativeness" of *must*-claims derives from the fact that they are, as it were, ear-marked for defeasibility. In using such a claim, a speaker points to the fact that the claim made might be overridden by later, more conclusive evidence.

The process of marking claims as defeasible can be overridden by conclusively establishing that the prejacent of the *must*-claim holds:

(21) The ball is in box A, B or C. It is not in box A. It is not in box B. So it must be in box C.

After processing the first three sentences, the final sentence merely states the obvious, marking that a conclusion has been drawn: so, in this example, *must* is reduced to the evidential signal diagnosed by von Fintel and Gillies.

6.2.2. *Must* vs. *Should*. Next, we consider the force relation between *must* and *should*. This matter is easier to decide, as the notion that *must* is stronger than *should* does not seem to be contested in the literature. So we can proceed much more quickly here.

The observation that *should* can be strengthened to *must*, but not vice versa, provides empirical evidence that the latter is indeed stronger than the former:

(22) Where are my diamonds?

a. They should be in the safe. In fact, they must be in the safe.

b. They must be in the safe. ?In fact, they should be in the safe.

That a stronger claim is made by saying that something must be the case than by saying that something should be the case is also plausible in view of the parallel to *deontic* modality. The following pair is taken from Silk (2012):

(23) a. I should help the poor. In fact, I must.

b. I must help the poor. ?In fact, I should.

In a recent paper, Krzyżanowska, Wenmackers, and Douven (2013) argue that *must* and *should* carry distinct evidential signals. The authors suggest that *must* signals the presence of an abductive inference, while *should* signals an inductive inference. Examples like the following two support this hypothesis:

(24) a. John has done many assignments well. He should be able to handle this one, too.

b. John has done many assignments well. ?He must be able to handle this one, too.

John's past performance provides the basis for an inductive inference. The fact that (24a) is considerably less marked compared to (24b) supports the position of Krzyżanowska et al. (2013) (note that 24b has a deontic reading on which it is quite acceptable). And conversely:

(25) a. Only John had access to the kitchen yesterday. He must have stolen the cookies.

b. Only John had access to the kitchen yesterday. ?He should have stolen the cookies.

The observation that access to the kitchen was restricted to John sets up an abductive inference to the extent that John is the one who took the cookies, as no other equally good explanation seems to be available.

Returning to the cookie example, we note that the conviction that John did it is easily defeated by taking further aspects into account:

(26) But perhaps the cookies were already stolen the day before.

This seems to me also the case where *must* is weak in the relevant sense: an information state that supports the claim that John must have stolen the cookies may easily evolve into a state in which this claim is no longer upheld.

Notice, further, that the question just which type of evidence gives naturally rise to *must* vs. *should* claims is independent of the claim that *must* carries stronger conviction than *should*. If Krzyżanowska et al. (2013)'s claim is correct, the conclusion to draw would be then that natural language treats abductive inferences as stronger than inductive ones.[11]

[11]It seems to me worth pondering whether the taxonomy used in Krzyżanowska et al. (2013) is comprehensive. Consider (Frank Veltman, p.c.):

(27) Normally, John does well on assignments. He should be able to pass this exam, too.

The inference from John's usual success to the particular case of this exam does not seem to be inductive. Rather, it seems to be a default inference (Veltman 1996).

Interestingly, Veltman (1996) saw a tight link between epistemic *must* and default reasoning. But there are some doubts whether the following discourse is really felicitous:

6.2.3. *Should* vs. *May*. That *should* is stronger than *may* almost goes without saying. The former is a "necessity modal", while the latter is a "possibility modal." An empirical test to decide to which class a given modal belongs is to check whether enumerating a number of options is acceptable with a particular modal:

(29) a. John might be in London and he might be in Paris.
b. John may be in London and he may be in Paris.
c. John could be in London and he could be in Paris.
d. ?John should be in London and he should be in Paris.
e. ?John must be in London and he must be in Paris.

As it turns out, claiming that two inconsistent facts should or must be the case is incoherent; but not so for *might*, *may and* could. This makes them qualify as "existential".

In view of this observation, the following data is not surprising:

(30) Recursion Theory is so difficult.
a. Don't worry. You should be able to pass it.
b. Don't worry. ?You may be able to pass it.

The second discourse above is marked, and the explanation seems to be that learning that one *may* be able to pass an exam is not bound to diffuse worries of failure. On the other hand, learning that one *should* be able to pass seems to instill at least some confidence in the hearer. That is: the *should*-claim gives rise to a higher "degree of acceptance" of the fact that one will pass the exam than the corresponding *may*-claim.

6.2.4. *May* vs. *Could* vs. *Might*. We turn to the trio of "existential modals": *may*, *could* and *might*. Here, the situation is again a bit more complicated. My own feeling is that *may* is stronger than *could*, which is in turn stronger than *might*. But finding empirical judgements that unequivocally support this position is difficult.

(28) Normally, John does well on assignments. ?He must be able to pass this exam, too.

So is perhaps *should* the epistemic modal that goes best with default reasoning? I will leave the matter unresolved.

The following is the best I can come up with. Which of the following stories is more scary?[12]

(31) a. A wolf might come in. It would eat you first.

b. A wolf could come in. It would eat you first.

c. A wolf may come in. It would eat you first.

My sense is that they are increasingly scary from (a.) to (c.). To me, (a.) seems to talk about an "abstract possibility", while (c.) points to a "real option". And (b.) lies somewhere inbetween. I draw the tentative conclusion that *may* is indeed stronger than *could*, which is in turn stronger than *might*. This account may need to be revised, which makes the following account a bit tentative.

Summarizing the preceding discussion, we arrive at the following picture:

$$\text{Is} < \text{Must} < \text{Should} < \text{May} < \text{Could} < \text{Might}$$

Under the assumption that the order $<$ is transitive, we conclude that our five modals lie on a linearly ordered strength scale, topped by categorical claims that something *is the case*.[13] So we have entailment from top to bottom in the following list:[14]

(32) a. John is in his office.

b. John must be in his office.

c. John should be in his office.

d. John may be in his office.

e. John could be in his office.

f. John might be in his office.

[12]Frank Veltman (p.c) tells me that the third discourse is infelicitous, and needs to be replaced by "A wolf may come in. It will eat you first." More empirical investigation seems needed to resolve this matter.

[13]With the proviso that, as admitted above, the evidence about the existential modals I have presented is rather sketchy.

[14]Not everyone is fully happy with this picture. Frank Veltman (p.c) thinks that, while *may* is stronger than *might*, the two modals *could* and *might* are really equally strong in their epistemic use. This would correspond to the following picture:

$$\text{Is} < \text{Must} < \text{Should} < \text{May} < \text{Could} \approx \text{Might}$$

I will have to leave the matter unresolved, pointing out once more the tentative flavour of my analysis of the existential modals.

6.3. *Acts of Acceptance*

We move towards a more technical account. The purpose of the previous section was to argue that the modal auxiliaries lie on a strength scale that is topped by categorical claims that something is the case. In this section, we try to extract constraints on the dynamic attitudes that could in principle be seen as capturing "acts of acceptance" in the relevant sense of providing a semantics for epistemic modal claims. Compared to the previous section, the approach is rather top-down in fashion, as it looks at what the auxiliary verbs we consider have in common. I will work with the simplifying assumption that our target class of five modals (plus the copula *is*) is representative for epistemic modals more generally. So we try to extrapolate properties of acts of acceptance from linguistic judgements about the members of our small class. We identify five such properties (strictness, triviality, affirmativity, defeasibility, and informativeness) and discuss to what extent they apply to *all* verbs we are considering, or just to some of them. Towards the end of the section, we evaluate what choices the analysis leaves open in terms of a suitable semantis for epistemic modals.

6.3.1. Strictness. A dynamic attitude τ is *strict* iff for any plausibility order $\mathcal{S}$, and proposition P:

$$\text{If } P \cap \mathrm{S} = \varnothing \text{ then } \mathcal{S}^{\tau P} = \varnothing.$$

More simply, but equivalently:

$$\mathcal{S}^{\tau \varnothing} = \varnothing.$$

This property is motivated by the observation that a positive claim, embedded under a modal or not, is unacceptable after its negation has been accepted:

(33) a. John is not in London. ?He is in London.

b. John is not in London. ?He must be in London.

c. ...

d. John is not in London. ?He might be in London.

6.3.2. Triviality. A dynamic attitude τ satisfies *triviality* iff for any plausibility order $\mathcal{S}$ and proposition P:

$$\text{If } P \cap \mathrm{S} = \mathrm{S} \text{ then } \mathcal{S}^{\tau P} = \mathcal{S}.$$

More simply, but equivalently:

$$\mathcal{S}^{\tau W} = \mathcal{S}.$$

The motivation for this property is that a positive claim, embedded under a modal or not, is trivial (uninformative) after that same positive claim has been accepted unembedded:

(34) a. John is in London. He is in London.

b. John is in London. He must be in London.

c. ...

d. John is in London. He might be in London.

6.3.3. Affirmativity. A dynamic attitude τ is *affirmative* iff for all plausibility orders $\mathcal{S}$, propositions P and worlds $w, v \in \mathrm{S}^{\tau P}$:

— if $w \in P$ and $v \notin P$, then $w \leq_{\mathcal{S}} v$ implies $w \leq_{\mathcal{S}^{\tau P}} v$, and
— if $w \notin P$ and $v \in P$, then $w \leq_{\mathcal{S}^{\tau P}} v$ implies $w \leq_{\mathcal{S}} v$, and
— if $w \in P$ iff $v \in P$, then $w \leq_{\mathcal{S}} v$ iff $w \leq_{\mathcal{S}^{\tau P}} v$.

The motivation for the first two clauses of this property is that after accepting a positive claim, embedded under a modal or not, an agent may come to regard it as more plausible that the prejacent is satisfied, but certainly not less.

(35) a. John is in London this weekend.

b. John should be in London this weekend.

c. John might be in London this weekend.

Accepting that John is in London this weekend, but at the same time finding it less plausible that John is in London this weekend before this very act of acceptance does seem to be extremely odd. This intuition seems to me to be deeply embedded into our understanding, not only of categorical claims, but of all epistemic modals under consideration. It also echoes the idea that the purpose of epistemic *might*-sentences is to "raise the possibility" of P, that has been advanced, for example, by Swanson (2006).

We have already met the third clause of the property under the name of conservation (cf. §3.7.1): accepting, for example, that John may be in London, does not give us any reason to re-evaluate the relative plausibility of two

worlds where John is in London, nor does it give us any reason to re-evaluate the relative plausibility of two worlds where John is *not* in London. So performing an upgrade τP really only comes down to affirming P, but nothing else.

6.3.4. Informativeness. A dynamic attitude τ is *informative* iff for any order $\mathcal{S}$ and $\mathcal{S}$-substantial proposition P: if $\neg P <_{\mathcal{S}} P$, then $\mathcal{S}^{\tau P} \neq \mathcal{S}$.[15]

Above, I have argued that epistemic modals (and the copula *is*) are subject to an affirmativity property. However, as it is formulated, the affirmativity of τ may be trivially satisfied. Notice that, for example, neutrality *id* (given by $\mathcal{S} \mapsto \mathcal{S}$ for any proposition P) is affirmative! But *id* is certainly not adequate to provide a semantics for any kind of modal! This is amended by informativeness.

Informativeness requires that, at least in a situation where all $\neg P$-worlds are strictly more plausible than all P-worlds, performing the upgrade τP will change the current plausibility order $\mathcal{S}$ in *some* way. In conjunction with the above affirmativeness property, this amounts to saying that epistemic modals (and *is*) raise the plausibility of their prejacent in a *non-trivial way*.

Should all epistemic modals under consideration satisfy this property? Veltman's *might* operator violates informativeness. We return to the issue below.

6.3.5. Defeasibility. A dynamic attitude τ is *defeasible* iff for any $\mathcal{S}$ and P: If $P \cap \mathrm{S} \neq \varnothing$, then $\mathrm{S}^{\tau P} = \mathrm{S}$.

Defeasibility of a dynamic attitude τ requires, essentially, any upgrade τP to provide soft information only, information that may be retracted as further information comes along (cf. Chapter §1 for more detail on the notion of "soft information"). Above, I have argued that epistemic *must* serves to mark its prejacent as representing defeasible evidence. In the previous section, our claim was that the epistemic modals under consideration are ordered by their modal force as follows:

$$\text{Is} < \text{Must} < \text{Should} < \text{May} < \text{Could} < \text{Might}$$

One thus expects that epistemic modals weaker than *must* are also defeasible. But should a claim that something *is* the case be construed as defeasible? The dynamic semantics tradition says no: traditionally, the interpretation of

[15] Recall the notation: given a plausibility order $\mathcal{S}$, we write $P <_{\mathcal{S}} Q$ iff for any $w, v \in \mathrm{S}$: if $w \in P$, $v \in Q$, then $w <_{\mathcal{S}} v$. Recall also from §3.2.4 that a proposition P is $\mathcal{S}$-substantial (in a plausibility order $\mathcal{S}$) if neither $P \cap \mathrm{S} = \varnothing$ nor $P \cap \mathrm{S} = \mathrm{S}$.

categorical claims like "John is in London" has been given in terms of world elimination, which amounts to a violation of defeasibility in the above sense. Again, we return to the issue below.

6.3.6. Summary. I have argued that all verbs under consideration should be interpreted by means of dynamic attitudes that are strict, satisfy triviality, and are affirmative; further, all the verbs which are strictly stronger than *might* should satisfy informativeness; finally, modals weaker than *must* (and including *must*) should receive a defeasible interpretation. And the question whether *might* should satisfy informativeness has been left open, as well as the question whether categorical claims could plausibly be construed as defeasible.

As for the last two points: I think how one decides on these questions depends largely on how one draws the line between semantics and pragmatics. I will adopt what I call the "classical view" here, but also comment on what I call the "pragmatic view", which seems promising to explore further.

6.3.7. The Classical View. Suppose one where only to commit to the position that acts of acceptance should satisfy strictness and triviality, but nothing more. Then we observe:

proposition 103. *For any dynamic attitude τ satisfying strictness and triviality:* $! \leq \tau \leq !^{\sim}$.

proof. Suppose that τ is satisfies strictness and triviality. We show first that $! \leq \tau$. To show this, we have to prove that $(\mathcal{S}^{!P})^{\tau P} = \mathcal{S}^{!P}$, for any plausibility order $\mathcal{S}$ and proposition P. But observing that $P \cap \mathrm{S}^{!P} = \mathrm{S}^{!P}$, by the triviality constraint, it follows that $(\mathcal{S}^{!P})^{\tau P} = \mathcal{S}^{!P}$. So our claim holds, and this shows that $! \leq \tau$.

Next, we show that $\tau \leq !^{\sim}$. To show this, we have to prove that $(\mathcal{S}^{\tau P})^{!^{\sim} P} = \mathcal{S}^{\tau P}$, for any plausibility order $\mathcal{S}$ and proposition P. We first observe that if $\mathcal{S}^{\tau P} = \varnothing$, then also $(\mathcal{S}^{\tau P})^{!P} = \varnothing$, and our claim holds. We may thus assume that $\mathcal{S}^{\tau P} \neq \varnothing$. By the fact that τ is strict, this implies that $\mathrm{S}^{\tau P} \cap P \neq \varnothing$. By definition of $!^{\sim}$, it follows that $(\mathcal{S}^{\tau P})^{!^{\sim} P} = \mathcal{S}^{\tau P}$, and, again, the claim holds. This shows that $\tau \leq !^{\sim}$. ⊣

Assuming we have not overlooked an act of acceptance that is stronger than the act of accepting a categorical claim, or an act of acceptance that is weaker than accepting that something might be the case, in view of the previous proposition it is but a small step to conclude that ! provides a reasonable

semantics for claims that something is the case, and $!^{\sim}$ provides a reasonable semantics for claims that something might be the case. This provides an underpinning for the "classical" position in update semantics: under our assumptions, ! and $!^{\sim}$ are the natural choices, and they are also the choices made in Veltman (1996).

6.3.8. The "Pragmatic" View. The pragmatic approach centers on two ideas: in using natural language to interact, we generally obtain information that is defeasible, adopting views that we may change as new evidence comes into view. Adopting such a view seems pertinent if one wants to analyze simple dialogues like the following:

(36) a. A: John is in London.

b. B: No, he might be in Paris, too.

Consider the perspective of a hearer witnessing this dialogue. The hearer trusts both speakers. So based on what speaker A says, the hearer first comes to accept that John is in London. But then, she comes to accept that John might be in Paris, too. So she revises her beliefs. To accomodate this sort of phenomena, interpreting acts of acceptance as defeasible, i.e., "up for revision" in general, seems useful.

The second idea is based on the intuition that *might* claims sometimes provide non-trivial, genuinely useable information. The following example is due to Paul Dekker:

(37) I told you it might rain!

Here, the speaker seems to point out that his past act of telling the hearer that it might rain was meant as a warning. And since warnings are typically meant to provide genuine information—how does that square with an account of *might* that construes it as generally uninformative?

But the question is, of course, what we mean by "informative" here. Our formal notion of informativeness defined above formalizes the notion in terms of *changing the relative plausibility hierarchy among worlds*. Is that really what *might* does?

(38) A wolf might come in.

Does a hearer who accepts that a wolf might come in perform some mental operation akin to advancing the plausibility of some worlds where a wolf

comes in? I think what is rather required of the hearer is to admit that the possibility of a wolf coming in is not beyond the conceivable.

Why then is *might* felt to be informative? An answer is suggested in the literature: a claim that it might be that P draws attention to the possibility that P (Groenendijk et al. 1996, Ciardelli, Groenendijk, and Roelofsen 2009). While this aspect is neither captured by our setting in general, nor by our (formal) notion of informativeness in particular, adapting the present setting in a way that could account for "attentive meaning" seems feasible, and in fact, combining the two would seem to be an interesting project for future research.

Regardless on how one decides on these matters, the question is of interest what dynamic attitudes fall out if one assumes not only affirmativeness, but also defeasibility and informativeness. Here, we observe:

PROPOSITION 104. *For any dynamic attitude τ satisfying affirmativeness, defeasibility and informativeness: $\Uparrow^{+} \leq \tau \leq \Uparrow^{\sim +}$.*

PROOF. Let τ be a genuine modal. We first show that $\Uparrow^{+} \leq \tau$. Let $\mathcal{S}$ be a plausibility order, and P a proposition. If $S \cap P = \varnothing$, it follows that $\mathcal{S}^{\Uparrow^{+}P} = \varnothing = \varnothing^{\tau P} = (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$, and our claim holds. So we may suppose that $S \cap P \neq \varnothing$. By definition of $\Uparrow^{+}$, we have $S^{\Uparrow P} = S$, and since τ is defeasible, $(S^{\Uparrow P})^{\tau P} = S^{\Uparrow P}$. It remains to be shown that for any $w, v \in S$: $(w, v) \in \mathcal{S}^{\Uparrow^{+}P}$ iff $(w, v) \in (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$. So let $w, v \in S$. If $w \in P$ iff $v \in P$, the claim holds since τ is affirmative. So suppose that $w \in P$, $v \notin P$. Then $(w, v) \in \mathcal{S}^{\Uparrow^{+}P}$ by definition of $\Uparrow^{+}$, and, again since τ is affirmative, $(w, v) \in (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$, so again, the claim holds. Finally, suppose that $w \notin P$, $v \in P$. Then $(w, v) \notin \mathcal{S}^{\Uparrow^{+}P}$ by definition of $\Uparrow^{+}$, and, again since τ is affirmative, $(w, v) \notin (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$, s, again, the claim holds. Hence we have shown that $(w, v) \in \mathcal{S}^{\Uparrow^{+}P}$ iff $(w, v) \in (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$, from which we conclude that $\mathcal{S}^{\Uparrow^{+}P} = (\mathcal{S}^{\Uparrow^{+}P})^{\tau P}$. It follows that $\Uparrow^{+} \leq \tau$.

As the second part of the proof, we show that $\tau \leq \Uparrow^{\sim +}$. Let $\mathcal{S}$ be a plausibility order, and P a proposition. If $S \cap P = \varnothing$, then $\mathcal{S}^{\tau P} = \varnothing$, since τ is strict. But $\varnothing^{\uparrow P} = \varnothing$, so our claim holds. We may thus suppose that $S \cap P \neq \varnothing$. Assuming that it is not the case that $\neg P <_{\mathcal{S}} P$, it follows that $\mathcal{S} \vDash Sb^{\sim} P$ (i.e., P is remotely plausible in $\mathcal{S}$), and since $\overline{\Uparrow^{\sim +}} = Sb^{\sim}$, our claim holds. Suppose, then that, $\neg P <_{\mathcal{S}} P$. By the fact that τ is informative, it follows that $\mathcal{S}^{\tau P} \neq \mathcal{S}$. Since τ is defeasible, $S^{\tau P} = S$. So there exists a pair $(w, v) \in S \times S$ such that $(w, v) \in \mathcal{S}^{\tau P}$, $(w, v) \notin \mathcal{S}$. Since τ is affirmative, it is impossible that (a) $w, v \in P$, (b) $w, v \in \neg P$, or (c) $w \in \neg P$, $v \in P$. Hence $w \in P$, $v \in \neg P$. But then, $\mathcal{S} \vDash Sb^{\sim} P$, hence $\mathcal{S}^{\Uparrow^{\sim +}} = \mathcal{S}$. ⊣

From the "pragmatic" point of view, one could take this result to support the claim that categorical claims should be semantically represented using $\Uparrow^{+}$,

while epistemic *might* claims should be represented using $\Uparrow^{\sim+}$. But notice that the preceding result is also useful from the point of view of the "classical" position, in that the "classical" theorist could take it as providing support for the claim that $\Uparrow^{+}$ should figure as providing a semantics for *must*, while $\Uparrow^{\sim+}$ should figure as providing a semantics for *may*!

Incidentally, this perspective promises to provide an explanation for the intuition voiced above that *may* is stronger than *might*: while *might*, at best, serves to draw attention to a possibility, existential modals stronger than *might* serve to raise the plausibility of the prejacent.

6.3.9. Semantics for Epistemic Modals. As announced above, we adopt what I have called the classical view here. Then, the following picture emerges:

1. ! (fixed point: K) — *is*
2. $\Uparrow^{+}$ (fixed point: Sb) — *must*
3. $\uparrow^{+}$ (fixed point: B) *should*
4. $\uparrow^{\sim+}$ (fixed point: $B^{\sim}$) — *may*
5. $\Uparrow^{\sim+}$ (fixed point: $Sb^{\sim}$) — *could*
6. $!^{\sim}$ (fixed point: $K^{\sim}$) — *might*

Of these, the dynamic attitudes in (1.)–(5.) satisfy informativeness, while the dynamic attitudes in (2.)–(6.) satisfy defeasibility. All dynamic attitudes in (1.)–(6.) satisfy strictness, triviality and affirmativity.

The choices made in (1.), (2.), (5.) and (6.) are motivated by Proposition 103 and Proposition 104 above, in conjunction with the observation that !, $\Uparrow^{+}$ and $!^{\sim}$ are canonical for their fixed point (Proposition 59), and $\Uparrow^{\sim+}$ is the unique dynamic attitude that is positionally optimal for its fixed point (Proposition 67). Choices (3.) and (4.) are motivated by the fact that *simple belief* B corresponds to the weakest natural form of (static) acceptance in our setting (essentially derived from the privileged position of the *most plausible worlds* in a plausibility order)—and the fixed point of $\uparrow^{+}$ is belief (Proposition 8); on the other hand $B^{\sim}$ is, as the dual of B, the strongest form of affirmation that falls short of proper acceptance. Note that both $\uparrow^{+}$ and $\uparrow^{\sim+}$ are also the unique dynamic attitudes that are positionally optimal for their fixed point (Proposition 67).

Conclusion

This dissertation has explored the concept of a dynamic attitude as a formal representation of an agent's assessment of the reliability of a source of information. Our formalization of dynamic attitudes (as functions that, given an informational input, map information states to information states) has drawn heavily on existing work in belief revision theory, dynamic epistemic logic and dynamic semantics. We have contributed a framework that has enabled us to explore new research directions. In conclusion, let me selectively highlight the main contributions of the dissertation, and point out some directions that might be taken in future work.

Contributions. The study carried out here has emphasized the importance of going beyond the discussion of specific belief revision policies, mainly in three respects. *First*, from the perspective adopted here, not only operations that induce belief are interesting objects of study. In this regard, our notion of a dynamic attitude generalizes the notion of a belief revision policy, and a wider range of interesting phenomena comes into view. In Chapter 2, we have seen that we can formalize notions of uniform trust, but also distrust, semi-trust, and mixed trust using our framework. *Second*, our notion of a fixed point, that was introduced in Chapter 1 and played a central role in the discussion of Chapter 2 and Chapter 3, embodies the idea that dynamic attitudes should be studied in tandem with the propositional attitudes they realize (an idea the roots of which go at least back to the research program outlined in van Benthem (1996)). Rather than arguing that a single belief revision policy is "the right one", the tandem approach leads to a more pluralistic perspective: depending on the particular target of revision, different policies may be adequate. This perspective has allowed us to clarify what is special about examples of dynamic attitudes that are well-known from the literature (they have natural fixed points! cf. §2.5), and, indeed, in which sense they are unique: Chapter 3 introduced the crucial notion of optimality (§3.2), according to which, roughly, a dynamic attitude is optimal if it realizes its fixed point in a way that adheres to the principle of minimal change. As it turns

out, infallible trust ! is the *unique* dynamic attitude that is optimal for irrecovable knowledge K (Corollary 59.1); strict strong trust $\Uparrow^+$ is the unique dynamic attitude that is optimal for strong belief Sb (Corollary 59.3); and strict minimal trust $\uparrow^+$ is the unique dynamic attitude that is positionally optimal for simple belief B (Proposition 66). These results, and others of a similar kind, are conceptually important, as they provide tight links between dynamic and propositional attitudes that are well-known from the literature.

Third, our aim has been to work towards results that are apt to provide insights about the landscape of dynamic attitudes as a whole. Several of the preceding chapters reported progress on that count.

In Chapter 2, we have considered the question which sources an agent needs to have at her disposal to be able to reach any information state from her current state by means of a sequence of upgrades. Theorem 28 and Theorem 29 provide some first answers, in terms of the classes of positive and semi-positive dynamic attitudes.

In Chapter 3 we asked just which dynamic attitudes are canonical, i.e., uniquely optimal for the fixed point they realize. Theorem 56 provided two criteria that are sufficient for canonicity.

Chapter 4 investigated the preservation properties of propositional attitudes. Theorem 74, Theorem 75 and Theorem 76 provide characterizations of the propositional attitudes that are preserved under substructures, respectively persistent, in terms of specific classes of dynamic attitudes, making crucial use of the notion of semi-distributivity, which derives from previous work in dynamic semantics.

In Chapter 5, we have introduced logical languages for dynamic attitudes. Proposition 92 and Theorem 96 provide versions of reduction and completeness theorems well-known from the dynamic epistemic logic literature that apply to all dynamic attitudes definable in the basic epistemic-doxastic language.

Future Directions. Our setting has a number of limitations that should be lifted eventually. A recent line of work in dynamic epistemic logic initiated by van Benthem and Pacuit (2011) takes non-total preorders as its starting point; in formal semantics, this seems to have been the dominant practice all along (cf., e.g., Kratzer (1981), Veltman (1996)). Considering dynamic attitudes on non-total preorders would bring the current setting closer to both. Furthermore, for reasons of generality, dynamic attitudes should be studied on infinite preorders, both well-founded and non-wellfounded ones. Finally, it would be interesting to know which dynamic attitudes can be defined by means of the action-priority operator introduced by Baltag and Smets (2008).

Going further, we mention a number of follow-up projects that naturally originate from the work presented here.

A characterization of the canonical propositional attitudes (defined as the propositional attitudes A for which there exists a unique optimal dynamic attitude τ such that the fixed point $\overline{\tau}$ of τ is A) in terms of an illuminating set of sufficient and necessary conditions remains to be found. The search may be combined with a more systematic exploration of alternative measures of similarity. §3.7, which focused on the case of simple belief, has merely scratched the surface in this direction.

The preservation results established in Chapter 4 are only the beginning of a wider-ranging investigation. We have already outlined a number of questions in this area, cf. the remarks after the proof of Theorem 76 in §4.2.

The notion of definability of dynamic attitudes could also form the starting point of a more extensive study. As we have pointed out in §5.2.4, the epistemic-doxastic language $\mathcal{L}$ we have considered (with operators for the two "knowledges": defeasible knowledge $\Box$ and irrecovable knowledge K) can define strong trust $\Uparrow$ and infallible trust $!$, but not minimal trust $\uparrow$. The question just which dynamic attitudes $\mathcal{L}$ *can* define is open. A more ambitious task is to classify different languages of epistemic logic by means of the dynamic attitudes they can define.

Another project is motivated by the notion of dynamic completeness introduced in §2.4.6. Precisely which sets of dynamic attitudes are dynamically complete? A more general (and more vague) question is: just which (types of) sources does an agent need to have at her disposal for her epistemic well-being? We make this question slightly more precise in the next paragraph, which discusses a theme that expands the scope of what we are trying to capture in our formal models.

Dynamics of Dynamic Attitudes. This thesis has explored the notion of a dynamic attitude in some detail, but has been silent on the dynamics of dynamic attitudes. We have explored the space between "informational stimulus" and "epistemic response", a space in which an agent chooses his response, i.e., decides how to change his mind based on the content of the stimulus, but also depending on how reliable he considers the source of information to be. But we have taken the result of the choice for granted, and have not investigated the circumstances in which an agent may choose to reconsider a choice made earlier. There is a question to be asked how an agent would come to adopt a particular dynamic attitude towards a particular source, and in view of what evidence he would change it. As Annette Baier observed, a trusting agent exposes his own vulnerability: "One leaves

others an opportunity to harm one when one trusts, and also shows one's confidence that they will not take it" (Baier 1986). So trust should not be bestowed upon one's sources too easily. But the agent who chooses to ignore all informational inputs, regardless of origin, cannot learn. So there are occasions where we need to give others the power to harm us, by trusting them, hoping that they won't.

Consequently, agents need to find a balance between "epistemic vigilance" (Sperber et al. 2010) and eagerness for information. Gullibility (trusting everyone) and exaggerated suspiciousness (trusting no one) are two extreme ways of making the trade-off, two examples of possible "meta-attitudes" guiding an agents' choices of dynamic attitudes, but one would like to know more about the options that lie in-between. Prima facie, it is not even clear what the criteria are for a reasonable mix of dynamic attitudes towards one's range of sources. A question one may ask is if an agent "too isolated" in the sense that he has "too few" sources at his disposal, or, on the contrary, "over-connected", in the sense of being exposed to "too many", potentially conflicting information sources. This is one direction in which the question posed at the end of the previous paragraph may be sharpened. The theme seems also related to recent application of epistemic logic to the theory of social networks (Seligman, Liu, and Girard 2013, Christoff and Hansen 2013).

Going further, an agent's general policy for assessing the trustworthiness of others needs to take into account the risk of epistemic wounds being inflicted. To indefinitely rely on those who hurt us is a recipe for disaster (in epistemic matters as elsewhere), but to permanently exclude potentially valuable sources of information from consideration may not be in our best interest either. The agent thus faces the need for revising his dynamic attitudes as new evidence comes in. He may choose to be vigilant until a source has proven worthy of trust; or he may lose trust as soon as he has obtained evidence for past deceit. He might change his attitude to a source once another, already trusted source vouches for the former, or discredits the former.

References

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

Horacio Arlo-Costa and Isaac Levi. Contraction: On the decision-theoretic origins of minimal change and entrenchment. *Synthese*, 152(1):129–154, 2006.

Guillaume Aucher. A combined system for update logic and belief revision. Master's thesis, ILLC, University of Amsterdam, 2003.

Annette C. Baier. Trust and anti-trust. *Ethics*, 96(2):231–256, 1986.

Alexandru Baltag and Lawrence Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Giacomo Bonanno, Wiebe van der Hoek, and Michael Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory*, volume 3 of *Texts in Logic and Games*. Amsterdam University Press, 2008.

Alexandru Baltag and Sonja Smets. Protocols for belief merge: Reaching agreement via communication. *Logic Journal of the IGPL*, in print.

Alexandru Baltag, Lawrence Moss, and Slawomir Solecki. The logic of public announcements, common knowledge and private suspicions. Technical report, CWI, 1999.

Alexandru Baltag, Hans van Ditmarsch, and Lawrence Moss. Epistemic logic and information update. In Pieter Adriaans and Johan van Benthem, editors, *Handbook of the Philosophy of Information*. Elsevier, Amsterdam, 2008.

Alexandru Baltag, Nina Gierasimczuk, and Sonja Smets. Belief revision as a truth-tracking process. In Krysztof Apt, editor, *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*. ACM, 2011.

Alexandru Baltag, Ben Rodenhäuser, and Sonja Smets. Doxastic attitudes as belief revision policies. In *Proceedings of the ESSLLI Workshop on Strategies for Learning, Belief Revision and Preference Change*. Opole, 2012.

Pierpaolo Battigalli and Marciano Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 105(2):356–391, 2002.

Johan van Benthem. Foundations of conditional logic. *Journal of Philosophical Logic*, 13(3):303–349, 1984.

Johan van Benthem. *Essays in Logical Semantics*. Reidel, Dordrecht, 1986.

Johan van Benthem. *Exploring Logical Dynamics*. CSLI Publications, Stanford, 1996.

Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14:129–155, 2007.

Johan van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2011.

Johan van Benthem and Maricarmen Martinez. The stories of logic and information. In Pieter Adriaans and Johan van Benthem, editors, *Handbook of the Philosophy of Information*. Elsevier, Amsterdam, 2008.

Johan van Benthem and Eric Pacuit. Dynamic logics of evidence-based beliefs. *Studia Logica*, 99(1):61–92, 2011.

Patrick Blackburn, Maarten de Rijke, and Yde Venema. *Modal Logic*. Cambridge University Press, 2001.

Oliver Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49:49–80, 2004.

Guido Boella and Leendert van der Torre. Normative multiagent systems and trust dynamics. In Rino Falcone, Suzanne Barber, Jordi Sabater-Mir, and Munindar P. Singh, editors, *Trusting Agents for Trusting Electronic Societies*, volume 3577 of *LNCS*, pages 1–17. Springer, Dordrecht, 2005.

Richard Booth and Thomas Meyer. Admissible and restrained revision. *Journal of Artificial Intelligence Research*, 26:127–151, 2006.

Craig Boutilier. Revision sequences and nested conditionals. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, pages 519–525, 1993.

Craig Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

Joan L. Bybee and Suzanne Fleischman. Modality in grammar and discourse: An introductory essay. In Joan L. Bybee and Suzanne Fleischman, editors, *Modality in Grammar and Discourse*, pages 1–14. Benjamins, New York, 1995.

Cristiano Castelfranchi and Rino Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In Yves Demazeau, editor, *Proceedings of the 3rd International Conference on Multi Agent Systems*, pages 72–79, 1998.

Zoé Christoff and Jens Ulrik Hansen. A two-tiered formalization of social influence. In Davide Grossi, Huaxin Huang, and Olivier Roy, editors, *Proceedings of the 4th International Workshop on Logic, Rationality and Interaction (LORI-4)*, volume 8196 of *LNCS*. 2013.

Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. Attention! Might in inquisitive semantics. In Ed Cormany, Satoshi Ito, and David Lutz, editors, *Proceedings of SALT XIX*. 2009.

Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1-2):1–29, 1996.

Mehdi Dastani, Andreas Herzig, Joris Hulstijn, and Leendert van der Torre. Inferring trust. In João Leite and Paolo Torroni, editors, *Proceedings of the 5th International Workshop on Computational Logic in Multi-Agent Systems*, pages 144–160, 2004.

Robert Demolombe. To trust information sources: A proposal for a modal logical framework. In Cristiano Castelfranchi and Yao-Hua Tan, editors, *Trust and Deception in Virtual Societies*, pages 205–225. Kluwer, Dordrecht, 2001.

Hans van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.

Hans van Ditmarsch and Barteld Kooi. The secret of my success. *Synthese*, 153:201–232, 2006.

Hans van Ditmarsch, Barteld Kooi, and Wiebe van der Hoek. *Dynamic Epistemic Logic*. Springer, Dordrecht, 2007.

Hans van Ditmarsch, Jan van Eijck, Floor Sietsma, and Yanjing Wang. On the logic of lying. In Jan van Eijck and Rineke Verbrugge, editors, *Games, Actions and Social Software*, number 7010 in LNCS, pages 41–72. Springer, Dordrecht, 2012.

Barbara Dunin-Kęplicz and Rineke Verbrugge. *Teamwork in Multi-Agent Systems: A Formal Approach*. Wiley and Sons, Chichester, 2010.

Jan van Eijck and Albert Visser. Dynamic semantics. *Stanford Encyclopedia of Philosophy*, 2008.

Rino Falcone and Cristiano Castelfranchi. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 740–747. ACM, 2004.

Kai von Fintel and Anthony S. Gillies. Must ... stay ... strong! *Natural Language Semantics*, 18(4):351–383, 2010.

Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge (Mass.), 1988.

Jelle Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, ILLC, University of Amsterdam, 1999.

Anthony S. Gillies. Iffiness. *Semantics and Pragmatics*, 3(4):1–42, 2010.

Jeroen Groenendijk, Martin Stokhof, and Frank Veltman. Coreference and modality. In Shalom Lappin, editor, *The Handbook of Contemporary Semantic Theory*, pages 179–213. Oxford Blackwell, 1996.

Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988.

Valentine Hacquard. Modality. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, pages 1484–1515. Mouton de Gruyter, Berlin, 2011.

Michael Halliday. Functional diversity in language as seen from a consideration of modality and mood in english. *Foundations of Language*, 6:322–361, 1970.

Joseph Halpern. *Reasoning about Uncertainty*. MIT Press, Cambridge (Mass.), 2003.

Gilbert Harman. *Change in View*. MIT Press, Cambridge (Mass.), 1986.

Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244, 2010.

Wilfried Hodges. *A Shorter Model Theory*. Cambridge University Press, 1997.

Wesley H. Holliday and Thomas Icard. Moorean phenomena in epistemic logic. In Lev Beklemishev, Valentin Goranko, and Valentin Shehtman, editors, *Advances in Modal Logic*, volume 8, pages 178–199. College Publications, London, 2010.

Richard Jeffrey. *Subjective Probability, The Real Thing*. Cambridge University Press, 2004.

Barteld Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–253, 2007.

Angelika Kratzer. The notional category of modality. In H.J. Eikmeyer and H. Rieser, editors, *Words, Worlds, and Context*, pages 38–74. 1981.

Angelika Kratzer. *Modals and Conditionals*. Oxford Studies in Theoretical Linguistics. Oxford University Press, 2012.

Karolina Krzyżanowska, Sylvia Wenmackers, and Igor Douven. Inferential conditionals and evidentiality. *Journal of Logic, Language and Information*, 22: 315–334, 2013.

Keith Lehrer. *Theory of Knowledge*. Westview Press, Boulder (Co.), 1990.

Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society*. Reidel, Dordrecht, 1981.

David Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.

Churn-Jung Liau. Belief, information acquisition and trust in multi-agent systems – a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.

Abhaya C. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41:353–390, 1994.

Frank Palmer. *Mood and Modality*. Cambridge University Press, 1990.

Anna Papafragou. Epistemic modality and truth conditions. *Lingua*, 116: 1688–1702, 2006.

Odile Papini. Iterated revision operations stemming from the history of an agent's observations. In Mary-Anne Williams and Hans Rott, editors, *Frontiers in Belief Revision*, pages 279–301. Kluwer, Dordrecht, 2001.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Stanley Peters and Dag Westerståhl. *Quantifiers in Language and Logic*. Clarendon Press, Oxford, 2006.

Jan Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic, and Z.W. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216. Oak Ridge National Laboratory, 1989.

John L. Pollock. Epistemic norms. *Synthese*, 71(1):61–95, 1987.

Ben Rodenhäuser. Dynamic attitudes, fixed points and minimal change. In Davide Grossi, Huaxin Huang, and Olivier Roy, editors, *Proceedings of the 4th International Workshop on Logic, Rationality and Interaction (LORI-4)*, volume 8196 of *LNCS*, pages 342–346. Springer, Heidelberg, 2013.

Daniel Rothschild and Seth Yalcin. On the dynamics of conversation. Unpublished Manuscript, 2012.

Hans Rott. Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, 61(2-3):469–493, 2004.

Hans Rott. Information structures in belief revision. In Pieter Adriaans and Johan van Benthem, editors, *Handbook of the Philosophy of Information*, pages 457–482. Elsevier, Amsterdam, 2008.

Hans Rott. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. In David Makinson, Jacek Malinowski, and Heinrich Wansing, editors, *Towards Mathematical Philosophy*, volume 28 of *Trends in Logic*, pages 269–296. Springer, Dordrecht, 2009.

Katrin Schulz. "If you'd wiggled A, then B would've changed": Causality and counterfactual conditionals. *Synthese*, 179(2):239–251, 2011.

Jeremy Seligman, Fenrong Liu, and Patrick Girard. Facebook and the epistemic logic of friendship. In Burkhard C. Schipper, editor, *Proceedings of the 14th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 229–238. ACM, 2013.

Alex Silk. Modality, weights and inconsistent premise sets. In Anca Chereches, editor, *Proceedings of SALT XXII*, pages 43–64. 2012.

Raymond Smullyan. *What is the Name of this Book? The Riddle of Dracula and other Logical Puzzles*. Prentice-Hall, New Jersey, 1978.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deidre Wilson. Epistemic vigilance. *Mind and Language*, 25(4):359–393, 2010.

Wolfgang Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In William L. Harper and Bryan Skyrms, editors, *Causation in Decision, Belief Change, and Statistics. Proceedings of the Irvine Conference on Probability and Causation*, volume II, pages 105–134. Kluwer, Dordrecht, 1988.

Wolfgang Spohn. A survey of ranking theory. In Franz Huber and Christoph Schmidt-Petri, editors, *Degrees of Belief*, pages 185–228. Springer, New York, 2009.

Robert Stalnaker. Assertion. *Syntax and Semantics*, 9:315–332, 1978.

Robert Stalnaker. *Inquiry*. Bradford, Oxford, 1984.

Robert Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128 (1):169–199, 2006.

Robert Stalnaker. Iterated belief revision. *Erkenntnis*, 70(2):189–209, 2009.

Eric Swanson. Something 'might' might mean. Unpublished Manuscript, 2006.

Frank Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25: 221–261, 1996.

Frank Veltman. Making counterfactual assumptions. *Journal of Semantics*, 22: 159–180, 2005.

Rineke Verbrugge and Lisette Mol. Learning to apply theory of mind. *Journal of Logic, Language and Information*, 17:489–511, 2008.

Douglas Walton and Erik Krabbe. *Commitment in Dialogue*. SUNY Press, Albany, 1995.

Gregory Ward, Betty J. Birner, and Jeffrey P. Kaplan. A pragmatic analysis of the epistemic *would* construction in English. In Roberta Facchinetti, Manfred Krug, and Frank Palmer, editors, *Modality in Contemporary English*. Mouton de Gruyter, Berlin, 2003.

Malte Willer. A remark on iffy oughts. *Journal of Philosophy*, 109(7):449–461, 2012.

Malte Willer. Dynamics of epistemic modality. *Philosophical Review*, 122(1): 45–92, 2013.

Malte Willer. An update on epistemic modals. *Journal of Philosophical Logic*, forthcoming.

Seth Yalcin. Non-factualism about epistemic modality. In Andy Egan and Brian Weatherson, editors, *Epistemic Modality*. Oxford University Press, 2011.

Abstract

While propositional attitudes—like knowledge and belief—capture an agent's opinion about a particular *piece* of information, dynamic attitudes, as understood in this dissertation, capture an agent's opinion about a particular *source* of information, more precisely: they represent the agent's assessment of (or opinion about) the reliability (or trustworthiness) of the source. The project of this dissertation is to study the latter notion from a general qualitative vantage point. The proposal of the thesis is to formally represent assessments of reliability by means of operations on information states: dynamic attitudes are encoded as strategies for belief change, capturing how an agent plans to "change her mind" once receiving a particular piece of information from a particular (type of) source. In this way, the dissertation establishes a connection to the rich existing literature on information dynamics, which has been a major focus of attention in belief revision theory, dynamic epistemic logic and dynamic semantics. The main focus of the work presented here is a study of the interplay between dynamic attitudes and the more well-known propositional attitudes.

In Chapter 1, we show that (introspective) propositional attitudes naturally arise as fixed points of dynamic attitudes; conversely, dynamic attitudes can be seen as chosen with a specific propositional attitude in mind which constitutes the target of belief change.

Chapter 2 studies various forms of trust and distrust, and intermediate forms of "semi-trust". More specifically, we identify a number of classes of dynamic attitudes that can be seen as capturing natural ways of assessing the reliability of a source, and typical representatives of each class. Also, we systematically relate them to the class of propositional attitudes using the notion of a fixed point.

Chapter 3 takes on the topic of minimal change that has traditionally played a foundational role in belief revision theory. The approach we suggest allows us to further study the question in which sense the typical dynamic attitudes identified in the previous chapter are really special (and, indeed, in many cases, canonical, that is, uniquely optimal for their fixed point).

In Chapter 4, we shift the perspective, and study the robustness (or preservation) of propositional attitudes under certain classes of transformations, devoting particular attention to preservation under substructures, a form of preservation that has traditionally been important in model theory.

Chapter 5 discusses the link between the static and the dynamic level that has received most attention in the dynamic epistemic logic literature. In this chapter, we study modal languages extended with dynamic modalities and show how the static base language can already define the dynamic part. This allows us to prove generic completeness theorems for our logics.

Chapter 6, finally, studies the formal setting developed here from another angle: we observe that the dynamic attitudes we have worked with so far can be interpreted not only as reliability assessments on behalf of an agent, but also as denotations for epistemic modals in natural language. Our main point is that the results of this dissertation are also of potential interest to the community working on the semantics of natural language.

Samenvatting

Propositionele attitudes, zoals kennis en geloof, verwijzen naar de opvattingen van een agent over een bepaald *stuk* informatie. Dynamische attitudes, zoals we het begrijpen in dit proefschrift, verwijzen naar de opvattingen over een bepaalde *bron* van informatie. Met andere woorden, deze dynamische attitudes duiden aan hoe een agent de betrouwbaarheid van de informatiebron inschat. Het plan in dit proefschrift is om deze notie van dynamische attitudes te bestuderen vanuit een algemeen kwalitatieve invalshoek. Het voorstel in dit werk is om zulke beoordelingen over de betrouwbaarheid formeel weer te geven door gebruik te maken van operaties op informatietoestanden. Dynamische attitudes hier worden beschreven als strategieën voor "belief change" die uitdrukken hoe een agent van plan is om haar opvattingen te wijzigen van zodra ze een specifiek stuk informatie van een bepaald (type) informatiebron ontvangt. Op deze wijze brengt dit proefschrift een verbinding tot stand met de bestaande literatuur over "information dynamics", literatuur die ook belangrijk is voor het werk in belief revision theory, dynamische epistemische logica en de dynamische semantiek. In dit werk wordt de aandacht helemaal gericht op de studie van de relatie tussen dynamische attitudes en de gekende propositionele attitudes.

In het hoofdstuk 1 tonen we aan dat (introspectieve) propositionele attitudes op natuurlijke wijze tot stand komen als de "fixed points" van dynamische attitudes. Anderzijds kunnen we dynamische attitudes ook zien als zijnde gekozen met een specifieke propositionele attitude in gedachten: deze vormt het doel van de geloofsverandering.

Het hoofdstuk 2 richt zich op de studie van verschillende vormen van "trust" en "disdrust", evenals tussenvormen van "semi-trust". Meer bepaald identificeren we een aantal klassen van dynamische attitudes en de typische vertegenwoordigers voor elke klasse, deze die gezien kunnen worden als natuurlijke wijzen om de betrouwbaarheid van een informatiebron te beoordelen. We verbinden dit dan op systematische wijze met de klassen van propositionele attitudes door gebruik te maken van de notie van "fixed point".

In hoofdstuk 3 gaan we in op het onderwerp van de "minimale verander-

ing" dat traditioneel ook binnen de belief revision theory een fundamentele rol speelde. De aanpak die we voorstellen laat toe om de vraag te bestuderen naar de wijze waarin de typische dynamische attitudes (die worden geïdentificeerd in de vorige hoofdstukken) echt speciaal zijn (en inderdaad, in veel gevallen, "canonical", i.e. uniek optimaal voor hun "fixed point").

In hoofdstuk 4 veranderen we het perspectief, we bestuderen de robuustheid (of het "behoud") van propositionele attidues onder bepaalde klassen van transformaties. Hierbij richten we onze aandacht op het behoud onder "substructures", een vorm van robuustheid die traditioneel van belang is binnen de model theorie.

Hoofdstuk 5 bespreekt de link tussen het statische en het dynamische niveau die binnen de dynamisch epistemische logica to nu toe de meeste aandacht heeft gekregen. In dit hoofdstuk bestuderen we modale talen, uitgebreid met dynamische modaliteiten en tonen we hoe de statische basistaal reeds het dynamisch deel kan definëren. Dit laat ons toe om volledigheidsstellingen te bewijzen voor onze logische systemen.

In het finale hoofdstuk 6, bestuderen we de formele setting vanuit een andere invalshoek: we merken op dat de dynamische attitudes waarmee we tot nu toe gewerkt hebben niet enkel geïnterpreteerd kunnen worden als "reliability assessments" voor een agent, maar ook als denotaties voor "epistemic modals" in de natuurlijke taal. Ons belangrijkste punt is dat de resultaten in dit proefschrift ook van potentieel belang zijn voor de gemeenschap die actief is in de semantiek van de natuurlijke taal.

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