A matter of trust: Dynamic attitudes in epistemic logic

Rodenhäuser, L.B.

Publication date
2014

Citation for published version (APA):
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 accommodating 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)](#). [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*). 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. 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. 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.

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 $\tau$, we shall take its target as given by the fixed point $\tau(\tau)$ of $\tau$, i.e., the propositional attitude realized by $\tau$ (cf. §1.7).

---

4Board (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.”


7A 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 $\tau$ “optimal” for its fixed point $\bar{\tau}$ if no other dynamic attitude realizes the same fixed point while changing given input orders less; and a dynamic attitude $\tau$ will be called “canonical” if it is uniquely optimal (i.e., $\tau$ 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.

Let $S$ and $S'$ be two (single-agent) plausibility orders. Thinking of $S'$ as the plausibility order resulting from $S$ 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 $S \Rightarrow S'$ (cf. §2.4.6 for the notation), i.e., there exists some upgrade $u$ such that $S^u = S'$, or, equivalently: $S' \subseteq S$.

We say that $S$ and $S'$ agree on $(w, v) \in S \times S'$ iff

$$(w, v) \in S \text{ iff } (w, v) \in S'.$$

The agreement set of $S$ and $S'$ is the set of pairs $(w, v)$ such that $S$ and $S'$ agree on $(w, v)$. Introducing notation:

$$\text{agree}_S S' := \{(w, v) \in S \times S' \mid S \text{ and } S' \text{ agree on } (w, v)\}.$$  

For any plausibility order $S$, the (strict) similarity order $(O_S, <_S)$ is then defined in the following way. First, $S' \in O_S$ iff $S' \subseteq S$ (i.e., iff $S \Rightarrow S'$). Second, for any $S', S'' \in O_S$:

$$S' <_S S'' \text{ iff } S'' \subset S' \text{ or } (S'' = S' \text{ and } \text{agree}_S S'' \subset \text{agree}_S S').$$

If $S' <_S S''$, then we say that $S'$ is more similar to $S$ than $S''$ (or that $S'$ is closer to $S$ than $S''$).

Notice that this definition has a “lexicographic” component: it favours keeping elements of the original domain $S$ over preserving pairs in the relation.

---

8We will return to the multi-agent setting only towards the end of Chapter 5.
$\leq_S$. Recall that $S$ represents the **hard information** of the agent, while $\leq_S$ 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 $S$, $S'$ and $S''$ are given as follows:

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

$$S' = \{(w), ((w, w))\},$$

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

Then $S'' \leq_S S'$, since $S' \subset S''$.

Going further, for any plausibility order $S$, the **weak similarity order** $(O_S, \leq_S)$ is defined by: $S' \leq_S S''$ iff either $S' <_S S''$ or $S' = S''$. More explicitly, this amounts to the following:

**Lemma 35.** $S' \leq_S S''$ iff either $S' \supset S''$ or $(S' = S''$ and agree$_S S' \supseteq$ agree$_S S'')$.

**Proof.** From left to right, $S' \leq_S S''$ implies by definition that $S' <_S S''$ or $S' = S''$. In the first case, $S' \supset S''$ or $S = S''$ and agree$_S S' \supset$ agree$_S S''$. In the second case, $S' = S''$ and agree$_S S' =$ agree$_S S''$. In either case, $S' \supset S''$ or $(S' = S''$ and agree$_S S' \supseteq$ agree$_S S'')$, which completes one half.

For the other half, suppose that $S' \supset S''$ or $(S' = S''$ and agree$_S S' \supseteq$ agree$_S S'')$. First, if $S' \supset S''$, then $S' <_S S''$, so $S' \leq_S S''$. Second, if $S' = S''$ and agree$_S S' \supset$ agree$_S S''$, then also $S' <_S S''$, so $S' \leq_S S''$.

Third, suppose that $S' = S''$ and agree$_S S' =$ agree$_S S''$. We claim that $S' = S''$, so $S' \leq_S S''$, which finishes the proof. To show the claim, consider any pair $(w, v) \in S' \times S' = S'' \times S''$. Suppose that $(w, v) \in$ agree$_S S'$. Then $(w, v) \in S'$ iff (by definition) $(w, v) \in S$ iff (by the assumption) $(w, v) \in S''$. On the other hand, suppose that $(w, v) \notin$ agree$_S S'$. If $(w, v) \in S'$, then (by definition) $(w, v) \notin S$, so (by the assumption) $(w, v) \in S''$; and analogously: if $(w, v) \notin S'$, then $(w, v) \notin S$, so $(w, v) \notin S''$. The claim holds. The proof is complete. $
$}

### 3.1.2. Properties of the Similarity Order

The weak similarity ordering $(O_S, \leq_S)$ has a number of desirable properties that are easy to check:

**Proposition 36.** $(O_S, \leq_S)$ is a partial order bounded by $S$ and $\emptyset$. In other words, for any $S', S'', S''' \in O_S$, the following properties are satisfied:

1. **Reflexivity:** $S' \leq_S S'$.
2. **Transitivity:** If $S' \leq_S S''$ and $S'' \leq_S S'''$, then $S' \leq_S S'''$.
3. **Antisymmetry:** If $S' \leq_S S''$ and $S'' \leq_S S'$, then $S' = S''$. 


4. Boundedness: If $S' \neq S$, then $S <_S S'$, and if $S' \neq \emptyset$, then $S' <_S \emptyset$.

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 $\tau$ optimal if it realizes its fixed point $\overline{\tau}$ in a way that is minimal compared to any other attitude $\sigma$ with the same fixed point $\overline{\tau}$. We cash this out by requiring that no such $\sigma$ is capable of making a "smaller step" along the similarity order. Formally, this yields the following notion of optimality.

3.2.1. Optimality. Let $\tau$ be a dynamic attitude. We say that $\tau$ is optimal iff there is no order $S$, proposition $P$ and dynamic attitude $\sigma$ such that

$$\overline{\sigma} = \overline{\tau} \text{ and } S^{\sigma P} <_S S^{\tau P}.$$  

Given a propositional attitude $A$, we say that $\tau$ is optimal for $A$ if $\tau$ 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 $\mathrel{\upharpoonright}$ is optimal.
3. Minimal trust $\mathrel{\upharpoonright}$ is optimal.
4. Neutrality id is optimal.
5. Isolation $\emptyset$ is optimal.

**Proof.**

1. Towards a contradiction, suppose that $!$ is not optimal. This implies that there exists a plausibility order $S$, a proposition $P$, and an attitude $\sigma$ such that $\overline{\sigma} = K = !$ and $S^{\sigma P} <_S S^{! P}$. The latter implies that either (1) $S^{\sigma P} \supset S^{! P}$ or (2) $S^{\sigma P} = S^{! P}$ and $\text{agree}_S S^{\sigma P} \supset \text{agree}_S S^{! P}$. Assuming (1), we observe that $P \cap S = S^{! P}$, and since $S^{\sigma P} \supset S^{! P}$, it follows that $S^{\sigma P} \cap \neg P = \emptyset$, hence $S^{\sigma P} \not\supset KP$, contradiction. Assuming (2), we derive a contradiction to the fact that $\text{agree}_S S^{! P} = S^{! P} \times S^{! P}$. 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 $\mathcal{P}$, and an attitude $\sigma$ such that $\mathcal{S} = \mathcal{B} \vee \neg \mathcal{K}$ and $\mathcal{S}^\mathcal{P} \prec_{\mathcal{S}} \mathcal{S}^{\mathcal{P}}$. If $P \cap \mathcal{S} = \emptyset$, then $\mathcal{S}^\mathcal{P} = \mathcal{S}^{\mathcal{P}} = \mathcal{S}$, contradiction (using the fourth item of Proposition 36). So $P \cap \mathcal{S} \neq \emptyset$. Also, $\mathcal{S}^\mathcal{P} = \mathcal{S}^{\mathcal{P}} = \mathcal{S}$, for otherwise, $\mathcal{S}^{\mathcal{P}} \prec_{\mathcal{S}} \mathcal{S}^{\mathcal{P}}$, contradiction. So either (1) there exist $w, v \in \mathcal{S}$ such that $w \preceq_{\mathcal{S}} v$, $w \preceq_{\mathcal{S}^\mathcal{P}} v$, but not $w \preceq_{\mathcal{S}^{\mathcal{P}}} v$, or (2) there exist $w, v \in \mathcal{S}$ such that not $w \preceq_{\mathcal{S}} v$, and not $w \preceq_{\mathcal{S}^\mathcal{P}} v$, but $w \preceq_{\mathcal{S}^{\mathcal{P}}} v$. 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 $\mathcal{S}^\mathcal{P} \not\equiv_{\mathcal{P}} \mathcal{K}$, we obtain a contradiction with the fact that $\mathcal{S}^\mathcal{P} = \mathcal{S}^{\mathcal{P}}$. We conclude with (2). From the fact that not $w \preceq_{\mathcal{S}} v$ and not $w \preceq_{\mathcal{S}^\mathcal{P}} v$, we infer that $w \in \mathcal{P}$, $v \notin \mathcal{P}$. Thus, from the fact that it is not the case that $w \preceq_{\mathcal{S}^\mathcal{P}} v$, we infer that $\mathcal{S}^{\mathcal{P}} \not\equiv_{\mathcal{P}} \mathcal{S}^\mathcal{P}$. But, again, we also have $\mathcal{S}^\mathcal{P} \not\equiv_{\mathcal{P}} \mathcal{K}$, so we have a contradiction to the initial assumption, and this concludes case (2). So $\uparrow$ is optimal for $\mathcal{B} \vee \neg \mathcal{K}$.

3. Towards a contradiction, suppose that $\downarrow$ is not optimal. Then there exists a plausibility order $\mathcal{S}$, a proposition $\mathcal{P}$, and an attitude $\sigma$ such that $\mathcal{S} = \mathcal{B} \vee \neg \mathcal{K}$ and $\mathcal{S}^\mathcal{P} \prec_{\mathcal{S}} \mathcal{S}^{\mathcal{P}}$. As in the proof of the previous item, we conclude that $P \cap \mathcal{S} \neq \emptyset$ and $\mathcal{S}^\mathcal{P} = \mathcal{S}$. So either (1) there exist $w, v \in \mathcal{S}$ such that $w \preceq_{\mathcal{S}} v$, $w \preceq_{\mathcal{S}^\mathcal{P}} v$, but not $w \preceq_{\mathcal{S}^{\mathcal{P}}} v$, or (2) there exist $w, v \in \mathcal{P}$ such that not $w \preceq_{\mathcal{S}} v$, and not $w \preceq_{\mathcal{S}^\mathcal{P}} v$, but $w \preceq_{\mathcal{S}^{\mathcal{P}}} v$. 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 $\downarrow$. It remains to consider case (d), i.e., we suppose $w \notin P, v \in P, w \preceq_{\mathcal{S}} v, w \preceq_{\mathcal{S}^\mathcal{P}} v$, while it is not the case that $w \preceq_{\mathcal{S}^{\mathcal{P}}} v$. We conclude by definition of $\downarrow$ that $v \in \text{best}_{\mathcal{S}} P$. We now make three observations:

- Since $\mathcal{S}^\mathcal{P} \equiv \mathcal{B} \mathcal{P}$ and $w \preceq_{\mathcal{S}^\mathcal{P}} v$, it follows that $v \notin \text{best}_{\mathcal{S}} \mathcal{P}$. So there exists $x \in \text{best}_{\mathcal{S}} \mathcal{P}$ such that $x \prec_{\mathcal{S}^\mathcal{P}} v$.
- Since $v \in \text{best}_{\mathcal{S}} P$, also $v \in \text{best}_{\mathcal{S}^{\mathcal{P}}} P$, hence $v \preceq_{\mathcal{S}^{\mathcal{P}}} x$.
- Since $v, x \in P$ and $v \in \text{best}_{\mathcal{S}} P$, it follows that $v \preceq_{\mathcal{S}} x$.

Overall, we now have the following situation: $v \preceq_{\mathcal{S}} x, v \preceq_{\mathcal{S}^{\mathcal{P}}} x$, while it is not the case that $v \preceq_{\mathcal{S}^{\mathcal{P}}} x$. This implies that $(v, x) \in \text{agree}_{\mathcal{S}^{\mathcal{P}}}$, $(v, x) \notin \text{agree}_{\mathcal{S}^\mathcal{P}}$. So it is not the case that $\mathcal{S}^\mathcal{P} \prec_{\mathcal{S}} \mathcal{S}^{\mathcal{P}}$, and this is a contradiction. Since both (1) and (2) yield a contradiction, $\downarrow$ is optimal for $\mathcal{B}$.

4. Immediate by the fact that $\mathcal{S} \equiv \top$ for any $\mathcal{S}$ and $\mathcal{S} \prec_{\mathcal{S}} \mathcal{S}'$ for any $\mathcal{S}'$ such that...
3.2. Optimality

\[ S \neq S' \text{ (cf. the fourth item of Proposition 36).} \]

5. Immediate by the fact that \( S \models \bot \) implies \( S = \emptyset \).

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

\[ \text{opt}_S AP := \{ S' \in O_S \mid S' \models AP, \exists S'' \in O_S : S'' <_S S', S'' \models AP \} . \]

Given an input order \( S \), the set \( \text{opt}_S AP \) captures the set of orders reachable from \( S \) that satisfy \( AP \) while differing minimally from \( S \). Notice that \( \text{opt}_S AP \) is, in general, non-empty:

**Lemma 38.** For any introspective propositional attitude \( A \), plausibility order \( S \), and proposition \( P \): the set \( \text{opt}_S AP \) is non-empty.

**Proof.** Choose an introspective propositional attitude \( A \), a plausibility order \( S \), and a proposition \( P \). We know from Theorem 5 that there is a dynamic attitude \( \tau \) such that \( \tau = A \). So there exists an order \( S' \in O_S \) such that \( S' \models AP \) and \( S <_S S' \), namely \( S' = S_\tau P \) (notice that \( S_\tau P \subseteq S \), as required for membership in \( O_S \)). It follows from the overall finiteness of our setting that \( <_S \) is well-founded. Thus, given that we have \( S' \), reachable from \( S \) and satisfying \( AP \), we will be able to find an order \( S'' \), reachable from \( S \), satisfying \( AP \), and being minimally different from \( S \). Hence \( \text{opt}_S AP \) is non-empty.

Notice also that if \( S \models AP \), then \( \text{opt}_S AP = \{ S \} \), since, by Proposition 36 (item 4), \( S <_S S' \) for any \( S' \neq S \).

3.2.4. Substantial Propositions. Let \( S \) be a plausibility order. \( P \) is insubstantial in \( S \) (or \( S \)-insubstantial) if \( S \cap P \in \{ \emptyset, S \} \). \( P \) is substantial in \( S \) (or \( S \)-substantial) otherwise, i.e., \( P \) is \( S \)-substantial iff \( \emptyset \subset S \cap P \subset S \) iff neither \( P \cap S = \emptyset \) nor \( P \cap S = S \).

As a consequence of Proposition 5 for any dynamic attitude \( \tau \), plausibility order \( S \) and proposition \( P \): if \( P \) is insubstantial in \( S \), then \( S_\tau P \in \{ S, \emptyset \} \). The next lemma will be useful in later sections.

**Lemma 39.** Suppose \( \sigma = \tau \). If \( P \) is \( S \)-insubstantial, then \( S_\sigma P = S_\tau P \).
Chapter 3. Minimal Change

Proof. Let $\mathcal{S}$ be a plausibility order, and suppose that $P$ is insubstantial in $\mathcal{S}$, i.e., $Q = P \cap S \in \{S, \emptyset\}$. Notice that $S^\tau P = S^\tau Q$ by dynamic conservativity (cf. §1.5.6). By Proposition 5, $S^\tau Q \in \{S, \emptyset\}$. Suppose first that $S^\tau P = S^\tau Q = S$. Then $S = \tau Q$, and since $\varphi = \tau$, we conclude that $S^\varphi Q = S$. Using dynamic conservativity again, $S^\varphi Q = S^\varphi P$. The claim holds. The second case is similar: suppose that $S^\tau P = S^\tau Q = S^\varphi$. Then $S^\varphi \varphi \tau Q$, and since $\sigma = \tau$, we conclude that $S^\sigma Q = S^\tau Q$. Using dynamic conservativity again, $S^\sigma Q = S^\sigma P$. So $S^\tau P = S^\tau Q$. The claim holds. So if $P$ is insubstantial in $\mathcal{S}$, then $S^\sigma P = S^\tau P$.

So any attitude $\tau$ that realizes a particular propositional attitude $A$ (in the sense that $\varphi = A$) realizes $A$ in exactly the same way for insubstantial propositions. For this reason, how $\tau$ treats insubstantial propositions has no weight for the question whether $\tau$ 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 $\tau$ with $\varphi = A$. 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 $\tau$ that is optimal for $A$.

Proof. The right to left direction is obvious: if $\varphi = A$, 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 $(S, P)$ such that $S$ is a plausibility order, $P \subseteq W$, and $P \cap S$ an element of $\text{opt}_S AP$. We now define a dynamic attitude $\tau$ as follows:

1. For any order $S$ and proposition $P$: if $S = AP$, then $S^\tau P := S$.
2. For any order $S$ and proposition $P$: if $S \neq AP$ then
   
   — if $P$ is insubstantial in $S$, then $S^\tau P := \emptyset$,
   
   — if $P$ is substantial in $S$, then $S^\tau P := f(S, P \cap S)$.

By the construction, $\tau$ is optimal for $A$. To show this, the main point is to notice that, given a plausibility order $S$ and a proposition $P$ such that $S \neq AP$ and $P$ is substantial in $S$, the function $f$ picks a plausibility order $f(S, P) \in \text{opt}_S AP$, which is to say that there exists no plausibility order $S'$ such that $S' <_S f(S, P)$ and $S' \neq AP$.
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 $\mathcal{K}$, the test for irrevocable knowledge, which is (recall §1.7.4) defined by

$$S^{?\mathcal{K}} := \begin{cases} S & S \models \mathcal{K}, \\ \emptyset & S \not\models \mathcal{K}. \end{cases}$$

**Proposition 41.** $\mathcal{K}$ is not optimal.

**Proof.** The fixed point of $\mathcal{K}$ is irrevocable knowledge $K$. To see why $\mathcal{K}$ is not optimal for $K$ (and thus not optimal), consider a plausibility order $S$ such that $S \cap P \neq \emptyset$ and $S \cap \neg P \neq \emptyset$, i.e., $S$ contains both $P$-worlds and non-$P$-worlds. By definition of $\mathcal{K}$: $S^{?\mathcal{K}} = \emptyset$. On the other hand, $S^P \neq \emptyset$, hence $S^P \supset S^{?\mathcal{K}}$, so $S^P \prec S^{?\mathcal{K}}$. Since $! = K$, it follows that $\mathcal{K}$ is not optimal. \(\rightarrow\)

The change induced by the upgrade $?KP$ in an order $S$ such that $S \not\models KP$ 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 $S$, 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 $*_n$ is the dynamic attitude which associates with each given plausibility order $S$ and proposition $P$ the plausibility order $S^{*nP}$ on the domain $S^{*nP} = S$, where for each world $w \in S$, the Spohn ordinal $\kappa_{S^{*nP}}(w)$ is given by

- if $P \cap S = \emptyset$, then $\kappa_{S^{*nP}}(w) = \kappa_S(w)$.
- if $P \cap S \neq \emptyset$, then
  - if $\kappa_S(P) > n$, then
    - if $w \in P$, then $\kappa_{S^{*nP}}(w) = \kappa_S(w) - n$
    - if $w \notin P$, then $\kappa_{S^{*nP}}(w) = \kappa_S(w) + n$
  - if $\kappa_S(P) \leq n$, then
    - if $w \in P$, then $\kappa_{S^{*nP}}(w) = \kappa_S(w) - (n - \kappa_S(P))$
Chapter 3. Minimal Change

— if \( w \notin P \), then \( \kappa_{S^{\uparrow n}P}(w) = \kappa_S(w) + (n - \kappa_S(P)) \)

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 \( *_nP \): 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 \( n \)th degree of trust \( \uparrow^n \) is the dynamic attitude which associates with each given plausibility order \( S \) and proposition \( P \) the plausibility order \( S^{\uparrow^n}P \) on the domain \( S^{\uparrow^n}P = S \), where for any \( w, v \in S \):

— if \( P \cap S = \emptyset \), then \( \kappa_{S^{\uparrow^n}P}(w) = \kappa_S(w) \).
— if \( P \cap S \neq \emptyset \), then
  — if \( \kappa_S(P) > n \), then
    — if \( w \in P \) and \( \kappa_{S|P}(w) \leq n \), then \( \kappa_{S^{\uparrow^n}P}(w) = \kappa_S(w) - n \)
    — if \( w \notin P \) or \( \kappa_{S|P}(w) > n \), then \( \kappa_{S^{\uparrow^n}P} = \kappa_S(w) + n \)
  — if \( \kappa_S(P) \leq n \), then
    — if \( w \in P \) and \( \kappa_{S|P}(w) \leq n \), then \( \kappa_{S^{\uparrow^n}P}(w) = \kappa_S(w) - (n - \kappa_S(P)) \)
    — if \( w \notin P \) or \( \kappa_{S|P}(w) > n \), then \( \kappa_{S^{\uparrow^n}P} = \kappa_S(w) + (n - \kappa_S(P)) \)

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 \( S \) which have Spohn rank greater than \( n \) in \( S|P \) are kept in place; but on performing an upgrade \( *_nP \), 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 \( *_nP \) is applied.
3.3. Non-Optimality

Nevertheless, $\uparrow^n$ and $\ast_n$ 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 point of the stricture of $\ast_n$ is the degree of belief $B^n$.

This observation makes $\uparrow^n$ and $\ast_n$ 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 $\ast_n$ 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 $S$, given by the spheres drawn in the figure. We show that $S^{\uparrow^2 P} \prec S^{\ast \ast P}$ (notice

**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 $\ast_2 P$ (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.
that the numbers in Figure 14.1 describe $S^{t2P}$, while the numbers in Figure 14.1 describe $S^{*2P}$.

We first notice that $S^{t2P} = S = S^{*2P}$. To show our claim, it is thus sufficient to establish that $\text{agree}_S S^{*2P} \subseteq \text{agree}_S S^{t2P}$, to which end we prove that $\text{agree}_S S^{*2P} \subseteq \text{agree}_S S^{t2P}$, while it is not the case that $\text{agree}_S S^{*2P} = \text{agree}_S S^{t2P}$.

To show that $\text{agree}_S S^{*2P} \subseteq \text{agree}_S S^{t2P}$, let $w, v \in S$. Observe:

- If $\kappa_S(w) \leq 3$ and $\kappa_S(v) \leq 3$, then $(w, v) \in \text{agree}_S S^{*1P}$ iff $(w, v) \in \text{agree}_S S^{t1P}$, and
- else $(w, v) \in \text{agree}_S S^{t1P}$.

These two observations are easy to check by inspecting Figure 14 and taken together they establish that $\text{agree}_S S^{*2P} \subseteq \text{agree}_S S^{t2P}$.

To show that it is not the case that $\text{agree}_S S^{*2P} = \text{agree}_S S^{t2P}$, take worlds $w, v \in S$ such that $\kappa_S(w) = \kappa_S(v) = 4$, $w \in P \cap S$, $v \in \neg P \cap S$. Then $(v, w) \in S$, $(v, w) \in S^{t2P}$, while $(v, w) \notin S^{*2P}$. So it is not the case that $\text{agree}_S S^{*2P} = \text{agree}_S S^{t2P}$. This yields the desired result: $S^{t2P} \not< S^{*2P}$, hence $S^{*2P}$ 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 $\tau$ is uniquely optimal in the sense that $\tau$ is the only optimal dynamic attitude whose fixed point is $\tau$. If this is the case, we will call $\tau$ canonical.

#### 3.4.1. Canonicity

Let $\tau$ be a dynamic attitude, and $A$ a propositional attitude.

- $\tau$ is canonical if $\tau$ is optimal and for any attitude $\sigma$: if $\overline{\sigma} = \overline{\tau}$ and $\sigma$ is optimal, then $\sigma = \tau$. 
— \( \tau \) is canonical for \( A \) if \( \tau \) 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* \(\emptyset\) 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 \( Sb \lor K\neg \); and “keep everything the same” to realize \( \tau \); and “delete everything” to realize \( \perp \).

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.
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 $S$ and $S$-substantial $P$, there exists an order $S' \in O_S$ such that $S' \models AP$ and for any order $S'' \in O_S$ such that $S' \not< S''$: if $S'' \models AP$, then $S' <_S S''$. 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 $S$, and an $S$-substantial proposition $P$, and find orders $S', S'' \in O_S$ such that $S' \models AP$, $S'' \models AP$, while $S'$ and $S''$ are incomparable in the relation $<_S$, that is: neither $S' <_S S''$, nor $S'' <_S S'$. 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 $S \models RP$ iff $w, v \in S : w \in P, v \not\in P \Rightarrow w \not\in P$
3.4. Canonicity

3. The disjunction of knowledge and opposite knowledge $K \lor K\neg$ is not canonical.

**Proof.** For each item, we supply an order $S$, a proposition $P$ substantial in $S$, and orders $S', S'' \in O_S$ such that neither $S' \prec_S S''$ nor $S'' \prec_S S'$. For the first item, cf. Figure 15; for the second item, cf. Figure 16; for the third item, cf. Figure 17.

v. The attitude $R$ should not be confused with refined belief $Rb$, as defined in §1.2.8.
To my mind, these three canonicity failures fall into two distinct categories. The fact that $R$ and $K \lor \neg K$ 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 $KP \lor K\neg P$. 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 $\tau$ is optimal, we need to compare it to arbitrary other dynamic attitudes with the same fixed point; to check whether $\tau$ 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 $\tau$ is actually $\tau$ 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
3.5. Characterizing Optimality and Canonicity 113

hard to see how such a characterization should look. Recall that a dynamic attitude \( \tau \) is optimal if there is no other dynamic attitude \( \sigma \) such that, given some proposition \( P \), there exists a plausibility order \( S \) such that \( S^{\sigma P} \) is more similar to \( S^{\tau P} \) than \( S^{\sigma P} \) (cf.§3.2.1). What we would like to see is that this is the case iff, given some plausibility order \( S \), and some proposition \( P \), applying \( \tau P \) picks in general a closest order \( S' \) satisfying \( S' \upharpoonright \tau P \); and canonical if it always picks the (unique) closest order \( S' \) such that \( S' \upharpoonright \tau P \).

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 \( \tau \) be a dynamic attitude. Let \( S_1, S_2 \) be plausibility orders such that \( S_2 \subseteq S_1 \). Suppose that \( P \) is \( S_1 \)-substantial. Assume that \( S_1 \upharpoonright \tau P, S_2 \upharpoonright \tau P \). Then there exists a dynamic attitude \( \sigma \) such that \( \sigma = \tau \) and \( S^{\sigma P} = S_2 \).

**Proof.** Define the dynamic attitude \( \sigma \) as follows:

\[
S^\sigma Q := \begin{cases} 
S_2 & S = S_1, Q \cap S = P \cap S, \\
S^{\tau Q} & \text{otherwise.}
\end{cases}
\]

Clearly, \( \sigma \) meets our requirements.

**Proposition 47.** Let \( \tau \) be a dynamic attitude. The following are equivalent:

1. \( \tau \) is optimal.
2. For any \( S \) and \( S \)-substantial \( P \): \( S^{\tau P} \in \text{opt}_S \tau P \).

**Proof.** From (1.) to (2.), we argue by contraposition. Suppose that there exists an order \( S \) and an \( S \)-substantial proposition \( P \) such that \( S^{\tau P} \notin \text{opt}_S \tau P \). Choose an element \( S' \in \text{opt}_S \tau P \) such that \( S' \prec_S S^{\tau P} \) (guaranteed to exist by Lemma 38; \( \text{opt}_S \tau P \) is non-empty). We observe that \( S \neq S' \), hence we know that \( S \notin \tau P \) and \( S' \neq \tau P \). So we may apply Lemma 46 to conclude that there exists a dynamic attitude \( \sigma \) such that \( S^{\sigma P} = S' \). Since \( S^{\sigma P} \prec_S S^{\tau P} \), it follows that \( \tau \) is not optimal.

From (2.) to (1.), we argue again by contraposition. Suppose that \( \tau \) is not optimal. Then there exists an order \( S \), a proposition \( P \), and an attitude \( \sigma \) such that \( \sigma = \tau \) and \( S^{\sigma P} \prec_S S^{\tau P} \). By Lemma 39, \( P \) is \( S \)-substantial (for otherwise, \( S^{\sigma P} = S^{\tau P} \), contradiction). Since \( S^{\sigma P} \upharpoonright \tau P \), it follows that \( S^{\tau P} \notin \text{opt}_S \tau P \), 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 \( \tau \): \( \tau \) is optimal iff \( \tau^+ \) is optimal.

**Proof.** Let \( \tau \) be a dynamic attitude. By Theorem 47, \( \tau \) is optimal iff for any plausibility order \( S \) and proposition \( P \) substantial in \( S \): \( S^{\tau P} \in \text{opt}_S \tau P \). But for any such \( S \) and \( P \): \( S^{\tau P} = S^{\tau^P} \) by definition of stricture (cf. §2.6.4, the last item in the list). So \( \tau \) is optimal iff \( \tau^+ \) is optimal.

Next, we extend our analysis to canonicity.

**Lemma 49.** Let \( \tau \) be an optimal dynamic attitude. Let \( S_1, S_2 \) be plausibility orders such that \( S_2 \subseteq S_1 \). Suppose that \( P \) is \( S_1 \)-substantial. Assume that \( S_1 \not\models \tau P, S_2 \models \tau P \). Assume that \( S_2 \in \text{opt}_{S_1} \tau P \). Then there exists an optimal dynamic attitude \( \sigma \) such that \( \sigma = \tau \) and \( S^\sigma P = S_2 \).

**Proof.** We define \( \sigma \) as in the proof of Lemma 46:

\[
S^\sigma Q := \begin{cases} 
S_2 & S = S_1, Q \cap S = P \cap S, \\
S^\tau Q & \text{otherwise.}
\end{cases}
\]

Clearly, \( \sigma \) meets our requirements.

**Proposition 50.** If \( \tau \) is canonical, then for any \( S \) and \( S \)-substantial \( P \): \( \text{opt}_S \tau P \) is a singleton set.

**Proof.** We show the contrapositive. Suppose that there exists an order \( S \), a proposition \( P \) such that \( P \) is substantial in \( S \), and suppose that \( \text{opt}_S \tau P \) is not a singleton. Since, generally, \( \text{opt}_S \tau P \neq \emptyset \) (Lemma 38), this means that \( \text{opt}_S \tau P \) has at least two elements.

Choose an element of \( \text{opt}_S \tau P \) such that \( S' \neq S^{\tau P} \) (guaranteed to exist by the previous step). By Lemma 49, there exists an optimal dynamic attitude \( \sigma \) such that \( \sigma = \tau \) and \( S^\sigma P = S' \). But \( \sigma \) is distinct from \( \tau \), since \( S^\sigma P \neq S^{\tau P} \). So \( \tau \) is not canonical. This shows our claim.

**Proposition 51.** Suppose that for any \( S \) and \( S \)-substantial \( P \): \( \text{opt}_S \tau P \) is a singleton. If \( \tau \) is optimal, then \( \tau \) is canonical.

**Proof.** Assume that for any order \( S \) and proposition \( P \) such that \( P \) is substantial in \( S \): \( \text{opt}_S \tau P \) is a singleton. Suppose that \( \tau \) is optimal. It follows
from Proposition\textsuperscript{47} that for any order $S$ and proposition $P$ substantial in $S$: $S^{\tau P} \in \text{opt}_S{\overline{\tau P}}$.

Suppose now that some dynamic attitude $\sigma$ is optimal for $\overline{\tau}$. Let $S$ be a plausibility order and $P$ a proposition. If $P$ is insubstantial in $S$, by Lemma\textsuperscript{39}, $S^{\sigma P} = S^{\tau P}$. If $P$ is substantial in $S$, by Proposition\textsuperscript{47}, using the fact that $\text{opt}_S{\overline{\tau P}}$ is a singleton, $S^{\sigma P} = S^{\tau P}$. So $\sigma = \tau$. By definition, $\tau$ 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 $\tau$ be a dynamic attitude. The following are equivalent:

1. $\tau$ is canonical.

2. For any $S$ and $S$-substantial $P$: $S^{\tau P} \in \text{opt}_S{\overline{\tau P}}$ and $|\text{opt}_S{\overline{\tau P}}| = 1$.

**Proof.** For the direction from (1.) to (2.), suppose that $\tau$ is canonical (and hence optimal). Let $S$ be a plausibility order, suppose that $P$ is substantial in $S$. By Proposition\textsuperscript{50}, the set $\text{opt}_S{\overline{\tau P}}$ is a singleton (one part of our claim). Since $\tau$ is optimal, by Proposition\textsuperscript{47}, $S^{\tau P} \in \text{opt}_S{\overline{\tau P}}$ (the second part of our claim). This shows the direction from (1.) to (2.).

For the converse direction, assume that the condition stated in (2.) holds. By Proposition\textsuperscript{47}, $\tau$ is optimal. By Proposition\textsuperscript{51}, $\tau$ is canonical. This shows the converse direction, and completes the proof.

The result says that a dynamic attitude $\tau$ 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: $\tau$ 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\textsuperscript{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\textsuperscript{54} characterizes the canonical propositional attitudes $A$ as those attitudes which allow us to find, for any
given order $S$ and $S$-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 $S$, proposition $P$ and introspective propositional attitude $A$: $\text{opt}_S AP = \text{opt}_S A(P \cap S)$.

**Proof.** Let $\Theta = \{S' \mid S \rightarrow S', S' \equiv A(P \cap S')\}$ by conservativity of $A$ (cf. §1.2.1). Observing that $(P \cap S') = (P \cap S) \cap S'$, it follows that $\Theta = \{S' \mid S \rightarrow S', S' \equiv A((P \cap S) \cap S')\}$. But then, again by conservativity of $A$, it follows that $\Theta = \{S' \mid S \rightarrow S', S' \equiv A(P \cap S)\}$. So $\text{opt}_S AP = \{S' \in \Theta \mid \nexists S'' \in \Theta : S'' <_S S'\} = \text{opt}_S A(P \cap S)$.

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 $S$ and $S$-substantial $P$: $|\text{opt}_S AP| = 1$.

**Proof.** From (1.) to (2.), suppose that $A$ is canonical. Then there exists a canonical dynamic attitude $\tau$ such that $\tau = A$. By Theorem 52, $|\text{opt}_S AP| = 1$ for any $S$ and $S$-substantial $P$.

From (2.) to (1.), suppose that for any $S$ and $S$-substantial $P$: $|\text{opt}_S AP| = 1$. Let $S$ be a plausibility order. Define a dynamic attitude $\tau$ as follows:

— If $P$ is insubstantial in $S$, put

$$S^{\tau P} := \begin{cases} S & S \equiv AP, \\ \emptyset & S \not\equiv AP. \end{cases}$$

— If $P$ is substantial in $S$, put

$$S^{\tau P} := S', \text{ where } S' \text{ is the unique element of the singleton } \text{opt}_S AP.$$
3.6. Canonical Propositional Attitudes

1. If \( P \in \{ W, \emptyset \} \), then \( P \) is insubstantial in \( S \), hence \( S^{\tau P} \in \{ S, \emptyset \} \) by definition of \( \tau \). So \( \tau \) satisfies strong informativity.

2. If \( P \) is insubstantial in \( S \), then \( S^{\tau P} \models AP \) by definition of \( \tau \). But if \( S^{\tau P} \models AP \), then \( (S^{\tau P})^{\tau P} = S^{\tau P} \), again by definition of \( \tau \). If, on the other hand, \( P \) is substantial in \( S \), then \( S^{\tau P} \in \text{opt}_S AP \) by definition of \( \tau \). But for any \( S' \in \text{opt}_S AP \): \( \text{opt}_{S'} AP = \{ S' \} \). Hence \( (S^{\tau P})^{\tau P} = S^{\tau P} \). So \( \tau \) is idempotent.

3. Notice that \( S \models AP \) iff \( S \models A(P \cap S) \) by conservativity of \( A \). By definition of \( \tau \), it follows that \( S^{\tau P} = S^{\tau(P \cap S)} \) for any order \( S \) and proposition \( P \) that is insubstantial in \( S \). Now suppose that \( P \) is substantial in \( S \). By Lemma \[53\] \( \text{opt}_S AP = \text{opt}_S A(P \cap S) \). It follows from the initial assumption that \( S^{\tau P} = S^{\tau(P \cap S)} \), using the definition of \( \tau \). So \( \tau \) satisfies dynamic conservativity.

We conclude that \( \tau \) is a dynamic attitude. But \( \tau = A \) by definition of \( \tau \). By Theorem \[52\] \( \tau \) 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 \( S \) and \( S \)-substantial \( P \), there exists an order \( S' \in O_S \) such that \( S' \models AP \) and for any order \( S'' \in O_S \) such that \( S' \neq S'' \): if \( S'' \models AP \), then \( S' <_S S'' \).

**Proof.** For the direction from (2.) to (1.), observe that from (2.), we easily infer that for any order \( S \) and \( S \)-substantial \( P \): \( |\text{opt}_S AP| = 1 \), 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 \( S \) and \( S \)-substantial \( P \): \( |\text{opt}_S AP| = 1 \). Take a plausibility order \( S \) and an \( S \)-substantial proposition \( P \), and let \( S' \) be the unique element of the singleton set \( |\text{opt}_S AP| = 1 \). Take any order \( S'' \in O_S \) distinct from \( S' \) satisfying \( S'' \models AP \); we claim that \( S' <_S S'' \). Suppose otherwise. Then there exists an order \( S''' \in \text{opt}_S AP \) such that \( S''' \neq S'' \). Clearly, \( S''' \neq S' \) (for otherwise, \( S' <_S S'' \), contradiction). It follows that \( |\text{opt}_S AP| \neq 1 \). This is a contradiction, so \( S' <_S S'' \), after all. This shows that (2.) holds. \( \neg \)}
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 \( S, S' \) be plausibility orders. \( S' \) is a refinement of \( S \) (or: \( S' \) refines \( S \)) if \( S = S' \) and \( S' \cup \{(w,v) \mid w \approx_S v\} = S \). We say that \( A \) is preserved under refinements iff for any plausibility orders \( S, S' \) and proposition \( P \): if \( S \models AP \) and \( S' \) is a refinement of \( S \), then \( S' \models AP \).

— Let \( S, S' \) be a plausibility order. \( S' \) embeds into \( S \) if \( S' \subseteq S \). We say that \( A \) admits joint embeddings iff for any plausibility orders \( S_1 \) and \( S_2 \) and proposition \( P \): if \( S_1 \models AP \) and \( S_2 \models AP \), then there exists an order \( S' \) such that \( S' = AP \) and \( S_1 \) and \( S_2 \) each embed into \( S' \), 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 \( S \) and propositions \( P \) substantial in \( S \), the following holds:

\[
\text{If } S \models S_1, S \models S_2, S_1 \models AP, \text{ and } S_2 \models AP, \text{ then } \exists S': S \models S', S' \preceq_S S_1, \\
S' \preceq_S S_2, \text{ and } S' \models AP.
\]

Then \( A \) is canonical.

**Proof.** Suppose that for all \( S \), for all substantial \( P \), the following holds: If \( S_1 \models AP, S_2 \models AP, \text{ and } S \models S_1, S \models S_2, \text{ then } \exists S': S \models S', S' \preceq_S S_1, S' \preceq_S S_2, \text{ and } S' \models AP. \)

Our aim is to show that \( A \) is canonical. For this purpose, let \( S \) be a plausibility order, and let \( P \) be a proposition substantial in \( S \). By Theorem \[54\] it is sufficient to establish that \( \text{opt}_S AP \) is a singleton set.

Proceeding to prove this, take two orders \( S_1, S_2 \in \text{opt}_S AP \). We show that \( S_1 = S_2 \), which proves our claim.

By the assumption, there exists \( S' \) such that \( S' \preceq_S S_1, S' \preceq_S S_2, S \models S', \text{ and } S' \models AP. \)
Since $S_1, S_2 \in \text{opt}_S AP$, it is not the case that $S' \prec_S S_1$, and it is not the case that $S' \prec_S S_2$. Hence $S' = S_1, S' = S_2$, and thus $S_1 = S_2$, 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 plausibility orders $S$ and propositions $P$ substantial in $S$:

If $S \rightarrow S_1, S \rightarrow S_2, S_1 \models AP$, and $S_2 \models AP$, then $\exists S': S \rightarrow S', S' \preceq_S S_1, S' \preceq_S S_2$, and $S' \models AP$.

**Proof.** Suppose that $A$ admits joint embeddings and $A$ is preserved under refinements.

Let $S_1$ and $S_2$ be plausibility orders such that $S_1 \models AP, S_2 \models AP, S \rightarrow S_1, S \rightarrow S_2$, with $P$ substantial in $S$.

We need to find a plausibility order $S'$ such that $S \rightarrow S', S' \preceq_S S_1, S' \preceq_S S_2$, and $S' \models AP$.

Our first observation is that, by the fact that $A$ admits joint embeddings, there exists a plausibility order $S_\#$ such that $S_\# \succeq S_1, S_\# \succeq S_2, S_\# = S_1 \cup S_2$ and $S_\# \models AP$. We observe that $S \rightarrow S_\#$.

We now prove an auxiliary observation. For any plausibility orders $S, T$ such that $S \rightarrow T$, let

$$agree^+_S T := \{(w, v) \in S \times S \mid (w, v) \in T, (w, v) \in S\},$$

$$agree^-_S T := \{(w, v) \in S \times S \mid (w, v) \not\in T, (w, v) \not\in S\}.$$  

We observe that, in general, $agree^-_S T = agree^+_S T \cup agree^+_S T$.

Returning to our orders $S_\#, S_1$ and $S_2$, we prove, for later use, the following claim:

**Claim.**

1. $agree^+_S S_\# \supseteq agree^+_S S_1 \cup agree^+_S S_2$.

2. If $(w, v) \in agree^-_S S_1 \cup agree^-_S S_2$ and $(w, v) \not\in agree^-_S S_\#$, then $w \equiv_{S_\#} v$.

**Proof.** For item (1), suppose that $(w, v) \in agree^+_S S_1$. Then $(w, v) \in S_1$, so $(w, v) \in S_\#$, and also $(w, v) \in S$, so $(w, v) \in agree^-_S S_\#$. Under the supposition that $(w, v) \in agree^-_S S_2$, make the analogous case. The claim follows.

For item (2), suppose that $(w, v) \in agree^-_S S_1, (w, v) \not\in agree^-_S S_\#$. Since $(w, v) \in agree^-_S S_1$, it follows that $(w, v) \not\in S_1, (w, v) \not\in S$. Since $(w, v) \not\in agree^-_S S_\#$, it follows that $(w, v) \in S_\#$. But since $(w, v) \not\in S_1$, it follows that $(v, v) \in S_1$, hence $(v, v) \in S_\#$. So $w \equiv_{S_\#} v$, the desired result.
We now return to the main thread, and use the plausibility order $S_h$ defined above to find the desired plausibility order $S'$.

Define $S' = (S', \leq_{S'})$ by means of setting $S' := S_1 \cup S_2$, and requiring, for all $w, v \in S'$:

- if $w \not\in S_h v$, then $w \leq_{S'} v$ iff $w \leq_{S_h} v$.
- if $w \in S_h v$, then $w \leq_{S'} v$ iff $w \leq_{S} v$.

We claim that $S'$ is a plausibility order. Reflexivity and connectedness are immediate by the fact that $S$ and $S_h$ are both reflexive and connected. To show that $S'$ is transitive, we argue as follows. Suppose that $(w, v), (v, x) \in S'$. We distinguish four possible cases, which we discuss in turn:

- Case 1: Suppose that $w \not\in S_h v$ and $v \not\in S_h x$. By the assumption, $w <_{S_h} v <_{S_h} x$, hence $w <_{S_h} x$. So $(w, x) \in S'$ by definition of $S'$.

- Case 2: Suppose that $w \not\in S_h v$ and $v \in S_h x$, i.e., $w \not\in S_h v \equiv S_h x$, and thus $w \equiv_{S_h} x$. So $(w, v), (x, v) \in S$ by the assumption, $(w, v) \in S$ by transitivity of $S$. So $(w, x) \in S'$ by definition of $S'$.

- Case 3: Suppose that $w \not\in S_h v$ and $v \in S_h x$. By the assumption, $w <_{S_h} v <_{S_h} x$, hence $w <_{S_h} x$. So $(w, x) \in S'$ by definition of $S'$.

- Case 4: Suppose that $w \in S_h v$ and $v \not\in S_h x$. Then, using the assumption, $w \equiv_{S_h} v <_{S_h} x$, hence $w <_{S_h} x$. So $(w, x) \in S'$ by definition of $S'$.

In each case, $(w, x) \in S'$, hence $S'$ is transitive, and thus indeed a plausibility order.

We now argue that $S'$ has all the properties we want. First, notice that $S'$ is a refinement of $S_h$, hence, by the assumption that $A$ is preserved under refinements, and recalling that $S_h \equiv AP$, we conclude that $S' \equiv AP$. Furthermore, $S \rightarrow S'$. It remains to be shown that $S' \preceq S S_1$ and $S' \preceq S S_2$.

To prove this, we show that $\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2$. This entails that $S' \preceq S S_1$ and $S' \preceq S S_2$, for the following reason: take $S_i$ with $i \in \{1, 2\}$. We know that $S' \supseteq S_i$. If $S' \supseteq S_j$, then $S' \preceq S S_j$, hence $S' \preceq S S_i$. If, on the other hand, $S' = S$, it is sufficient to show that $\text{agree}_S S' \supseteq \text{agree}_S S_1$ to be able to conclude that $S' \preceq S S_i$. Overall, this implies that if we are able to show that $\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2$, we are done.

This, then, is our final claim: $\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2$. Proceeding to show it, recall from above that $\text{agree}_S S_h \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2$ (this was item (1.) of the claim proven above). Notice that our definition of $S'$ ensures that $\text{agree}_S S_h \supseteq \text{agree}_S S'$: to obtain $S'$, we have exclusively deleted pairs $(w, v)$ from $S_h$ that are not in $S$, in other words: for any pair $(w, v) \in S_h$ such that $(w, v) \in S$, we have $(w, v) \in S'$. It follows that $\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2$. 

Next, suppose that \((w, v) \in \text{agree}_S S_i\), for some \(i \in \{1, 2\}\). This means that \((w, v) \notin S\). Suppose first that \((w, v) \notin S\), and by definition of \(S'\), \((w, v) \notin S'\), so \((w, v) \in \text{agree}_S S'\). Suppose, second, that \((w, v) \notin \text{agree}_S S'\). As we have seen above (item \((2.)\) of the claim proven above), this implies that \(w \approx_{S} v\). By definition of \(S'\), it follows that \((w, v) \notin S'\), hence, again, \((w, v) \in \text{agree}_S S'\). It follows that \(\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2\).

At this point, we are in a position to conclude that \(\text{agree}_S S' \supseteq \text{agree}_S S_1 \cup \text{agree}_S S_2\), and the proof is complete.

This also completes the proof of Theorem \ref{theo:canonicity:predicate}.

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 \(\tau\) which is optimal for \(A\) is sufficient to ensure that \(\tau\) 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 \(Sb \lor K\).\neg

3. *Strong positive trust \(\uparrow^+\)* is canonical for strong belief \(Sb\).

4. *Bare semi-trust \(\uparrow^*\)* is canonical for dual knowledge \(K\).

5. *Neutrality id* is canonical for triviality \(\top\).

6. *Isolation \(\emptyset\)* 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 \ref{theo:canonicity:predicate}, it follows that \(A\) is canonical. But if \(A\) is canonical, and \(\tau\) is optimal for \(A\), then \(\tau\) is canonical for \(A\). So we merely need to verify, for each item, that the \(\tau\) 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 \ref{prop:optimality:symmetric}. For item \((3.)\), the claim follows from Proposition \ref{prop:optimality:symmetric} and the fact that, by Corollary \ref{cor:optimality:symmetric}, optimality is preserved under strictures.

Noticing that optimality is invariant under opposites, one obtains analogous results for the opposites of the attitudes mentioned in Corollary \ref{cor:canonicity:predicate}.
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 constitutes 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 $P \cap S$ within a plausibility order $S$. 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 $S'$ that keeps the relative plausibility of $P$-worlds the same will count as more similar to the input order $S$ than an order $S''$ 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 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 $\tau$ is conserving iff for any order $S$, proposition $P$ and worlds $w, v \in S^P$:

$$if \ w \in P \ iff \ v \in P, \ then \ w \leq_S v \ iff \ w \leq_{S^P} v.$$ 

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. 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 $S$ be a plausibility order, suppose that $P$ is substantial in $S$, and let $S' \in \text{opt}_S BP$. The following hold:

1. $S' = S$.

2. For all $w, v \in S$: if $(w, v) \notin \text{best} S'$, then $(w, v) \in S' \iff (w, v) \in S$.

**Proof.** Let $S$ be a plausibility order, suppose that $P$ is substantial in $S$, and let $S' \in \text{opt}_S BP$. We consider the two items in turn.

1. Notice that $P \cap S \neq \emptyset$, since $P$ is, by assumption, substantial in $S$. Now we know that $S' \subseteq S$. But if $S' \subset S$, then $S^P \prec_S S'$ (since $S^P = S$), contradicting the assumption that $S' \in \text{opt}_S BP$. So $S' = S$.

2. We define the order $S''$ as follows:

- $X_1 := \{(y, z) \in S' \mid y \in \text{best} S'\}$,
- $X_2 := \{(y, z) \in S \mid y, z \notin \text{best} S'\}$,

Cf. Darwiche and Pearl [1996].
Proposition 61. Suppose that $\tau$ is conserving. If $\tau$ is optimal for belief, then $\tau = \tau^+$.

Proof. Suppose that $\tau$ is conserving and optimal for belief. Let $S$ be a plausibility order, and let $P \in W$. We discuss two cases. Suppose first that $P$ is insubstantial in $S$. Then $S^{\tau P} = S^{\tau^+ P}$, since $\tau = \tau^+$, so our claim holds. Suppose, second, that $P$ is substantial in $S$. Since $\tau$ is optimal, it follows by Proposition 47 that $S^{\tau P} \in \text{opt}_S BP$. 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 $w, v \notin \text{best}S^{\tau P}$, then $(w, v) \in S^{\tau P}$ if $(w, v) \in S$.

3. Since $\tau = \tau^+$, it follows that $\text{best}S^{\tau P} \subseteq P$. And since $\tau$ is conserving, it follows that for all $w, v \in S \cap P$: $w \leq_S v$ iff $w \leq_S v$, hence $\text{best}S^{\tau P} = \text{best}_S P$.

Combining these observations, we conclude that $S^{\tau P} = S^{\tau^+ P}$. So, again, our claim holds. This shows that $\tau = \tau^+$ 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:

$$\forall w, v \in S^{\tau P} : \text{if } w, v \in P, \text{ then } w \leq_S v \text{ iff } w \leq_{S^{\tau P}} v.$$  

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. The Case of Simple Belief

3.7.2. Weighted Similarity. For any proposition $P$ and plausibility orders $S$ and $S'$, let $\text{agree}_P(S, S') := \text{agree}(S, S') \cap (P \times P)$. Given a plausibility order $S$, and a proposition $P$, we define the order $(O_S, \prec_S^P)$ as follows:

$$S' \prec_S^P S'' \iff$$

- $S' \supset S''$, or
- $S' = S''$ and $\text{agree}_S^P S' \supset \text{agree}_S^P S''$, or
- $S' = S''$ and $\text{agree}_S^P S' = \text{agree}_S^P S''$ and $\text{agree}_S S' \supset \text{agree}_S S''$.

In comparing the similarity of two orders $S'$ and $S''$ to a given order $S$, 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 $\tau$ weighted optimal if there exists no attitude $\sigma$ and plausibility order $S$ such that $\sigma = \tau$ and $S^\sigma P \prec_S^{\tau P} S^{\tau P}$; and call $\tau$ weighted optimal for $A$ if $\tau$ is weighted optimal and $\tau = A$.

**Proposition 62.** If $\tau$ is weighted optimal for simple belief $B$, then $\tau = \uparrow^+$.

**Proof.** Suppose that $\tau$ is weighted optimal for simple belief. Let $S$ be a plausibility order, and $P$ a proposition. If $P$ is insubstantial in $S$, then $S^\tau P = S^{\uparrow^+ P}$. Suppose now that $P$ is substantial in $S$. We have three observations to make:

1. $S^\tau P = S$ (for otherwise, $S^{\uparrow^+ P} \prec_S^P S^{\tau P}$, contradiction).

2. $\text{best}_S S^\tau P = \text{best}_S P$. This follows from the fact that $S^{\tau P} = BP$, together with the fact that for any $w, v \in P \cap S$: $w \leq_{S^\tau P} v$ iff $w \leq_{S^{\uparrow^+ P}} v$ (for otherwise, $S^{\uparrow^+ P} \prec_S^P S^{\tau P}$, contradiction).

3. For all $w, v \in S$ such that $w, v \notin \text{best}_S S^\tau P$: $w \leq_{S^\tau P} v$ iff $w \leq_S v$ (for otherwise, $S^{\uparrow^+ P} \prec_S^P S^{\tau P}$, contradiction).

Taken together, these observations imply that $S^\tau P = S^{\uparrow^+ P}$. So regardless of whether $P$ is substantial or insubstantial in $S$, we have shown that $S^\tau P = S^{\uparrow^+ P}$. 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 $S', S'' \in O_S$, the question which of the two is “more similar” to $S$ 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 \( S' \) compared to an (“initial”) order \( S \) if, in going from \( S \) to \( S' \), 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 \( S' \) as it used to be in \( S \).

Formally, let \( S, S' \) be plausibility orders such that \( S \to S' \), and let \( w \in S' \). We define the proposition \( \text{maintain}_S S' \) by stipulating that \( w \in \text{maintain}_S S' \) iff for all \( v \in S' \):

- If \( v \approx_{S'} w \), then \( v \approx_S w \)
- If \( v <_{S'} v \), then \( v <_S w \).

Note that membership of \( w \) in \( \text{maintain}_S S' \) 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 \( \circ \), given an order \( S \), putting \( S' \circ S' \) iff:

- \( S'' \subset S' \), or
- \( S'' = S' \) and \( \text{maintain}_S S'' \subset \text{maintain}_S S' \) or
- \( S'' = S' \) and \( \text{maintain}_S S'' = \text{maintain}_S S' \) and \( \text{agree}_S S'' \subset \text{agree}_S S' \).

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 \( S' \circ S'' \), since \( S' = S'' \) and \( \text{maintain}_S S'' = \{y\} \subset \{x, y\} = \text{maintain}_S S' \).

Adopting now the notation of Figure 19, we notice again that \( S' \circ S'' \), since \( S' = S'' \) and \( \text{maintain}_S S'' = \{y\} \subset \{x, y\} = \text{maintain}_S S' \), for the same reason: \( \text{maintain}_S S'' = \{y\} \subset \{x, y\} = \text{maintain}_S S' \).

The reader may notice that, under the assumption that \( x, y \in P \) and \( w \notin P \), \( S' \) actually equals \( S^{\uparrow P} \), 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 below, while Proposition 66 records the desired result.

**Lemma 63.** Let \( S, S' \) be a plausibility order, let \( P \) be substantial in \( S \) and suppose that \( S' \equiv BP \). Then:

1. \( \text{maintain}_S S'^P = \text{best}_S P \cup \{v \in S | \exists w \in \text{best}_S P : w <_S v\} \).
The Case of Simple Belief

2. \( \text{maintain}_S S' \subseteq \text{maintain}_S S^T_P \).

3. If \( S' \in \text{opt}_S^\circ BP \), then \( \text{maintain}_S S' = \text{maintain}_S S^T_P \).

proof. The first item is immediate by definition of \( \uparrow \). As for the second item, towards a contradiction, suppose that \( \text{maintain}_S S' \notin \text{maintain}_S S^T_P \). Then there exists \( x \in \text{maintain}_S S' \) such that \( x \notin \text{maintain}_S S^T_P \). By the first item, \( x \preceq_S y \) for any \( y \in P \cap S \). So \( x \notin P \). Since \( x \in \text{maintain}_S S' \), it follows that for any \( y \in P \cap S \): \( x \preceq_S y \) (for otherwise, \( y \preceq_S x \), contradiction). It follows that \( \text{best}_S S' \notin P \), so \( S' \notin BP \). This contradicts our initial assumption, and the claim follows, which finishes the second item. For the third item, suppose that \( S' \in \text{opt}_S^\circ BP \). Notice that if \( \text{maintain}_S S' \subseteq \text{maintain}_S S^T_P \), it follows that \( S^T_P \preceq_S S' \), which contradicts the assumption that \( S' \in \text{opt}_S^\circ BP \). Hence, by the
conclude that, actually, agree.

**Lemma 64.** Let \( P \) be substantial in \( S \), suppose that \( S \rightarrow S' \), and let \( S' \in \text{opt}^2_S BP \).

1. \( S' = S \).
2. \( \text{best}_S S' = \text{best}_S P \).

**Proof.** Let \( P \) be substantial in \( S \), and let \( S' \in \text{opt}^2_S BP \). We consider the three items in turn.

1. We know that \( S' \subseteq S \). Suppose that \( S' \subset S \). Then \( S'^P \not\subseteq S' \), so \( S' \notin \text{opt}^2_S BP \), contradiction. Hence \( S' = S \).

2. We consider the two halves of the claim in turn and show them by reductio. For one half, suppose that \( \text{best}_S P \notin \text{best}_S S' \). Choose a world \( x \in S \) such that \( x \notin \text{best}_S P \), \( x \notin \text{best}_S S' \). Since \( S' = BP \), we have that \( \text{best}_S S' \subseteq P \), so there exists \( w \in P \cap S \): \( y <_S x \). Since \( x \in \text{best}_S P \), it follows that \( x \leq_S y \). So \( x \notin \text{maintain}_S S' \). By the previous lemma (first item), \( x \in \text{maintain}_S S'^P \). So \( \text{maintain}_S S' \neq \text{maintain}_S S'^P \). However, again by the previous lemma (third item), \( \text{maintain}_S S' = \text{maintain}_S S'^P \). This is a contradiction. We conclude that \( \text{best}_S S' \subseteq \text{best}_S P \).

For the other half, suppose that \( \text{best}_S P \notin \text{best}_S S' \). Choose a world \( x \in S \) such that \( x \notin \text{best}_S S' \), \( x \notin \text{best}_S P \). Since \( S' = BP \), it follows that \( x \in S \cap P \). Since \( x \notin \text{best}_S P \), there exists a world \( y \in \text{best}_S P \) such that \( y <_S x \). Since \( x \in \text{best}_S S' \), we have \( x \leq_S y \). So \( y \notin \text{maintain}_S S' \). Now we argue as in the first half of the proof for this item, and arrive at the desired result: \( \text{best}_S P \subseteq \text{best}_S S' \).

**Lemma 65.** Let \( P \) be substantial in \( S \), suppose that \( S \rightarrow S' \), and let \( S' \in \text{opt}^2_S BP \). Then \( S' = S'^P \).

**Proof.** We first observe that, since \( S' = S'^P = S \) (first item of Lemma 64) and \( \text{maintain}_S S' = \text{maintain}_S S'^P \) (third item of Lemma 65), the following holds by definition of \( <^0_S \):

\[
\text{If agree}_S S' \subset \text{agree}_S S'^P \ \text{then} \ S'^P <^0_S S'.
\]

It is thus sufficient to show that \( \text{agree}_S S' \subseteq \text{agree}_S S'^P \), from which we may conclude that, actually, \( \text{agree}_S S' = \text{agree}_S S'^P \) (for otherwise, we obtain a contradiction to the assumption that \( S' \in \text{opt}^2_S BP \)). But from \( \text{agree}_S S' = \text{agree}_S S'^P \) it follows that \( S' = S'^P \).

To show the claim, we establish that for any \( (w, v) \in S \times S \): if \( (w, v) \in \text{agree}_S S' \), then \( (w, v) \in \text{agree}_S S'^P \). Letting \( (w, v) \in S \times S \), we discuss two cases.
**The Case of Simple Belief**

**Case 1:** If at least one of \(w\) and \(v\) is an element of best \(S'\), our claim holds, for in that case, \((w, v) \in \text{agree}_S S'\) iff \((w, v) \in \text{agree}_S S^{\uparrow P}\) using the definition of \(\uparrow\) and the fact that best \(S' = \text{best}
\(S^{\uparrow P}\) (Lemma 64). **Case 2:** Suppose that \(w, v \not\in \text{best}
S'\). By definition of \(\uparrow\), \((w, v) \in \text{agree}_S S'\). Thus if \((w, v) \in \text{agree}_S S'\), then \((w, v) \in \text{agree}_S S^{\uparrow P}\).

This is the desired result, so we may, indeed, conclude that agree\(_S S' = agree_S S^{\uparrow P}\), and thus \(S' = S^{\uparrow P}\).

**Proposition 66.** Let \(\tau\) be a dynamic attitude and suppose that \(\tau\) is positionally optimal for belief. Then \(\tau = \uparrow^+\).

**Proof.** Let \(S\) be a plausibility order, and \(P \subseteq W\). If \(P\) is insubstantial in \(S\), then \(S^{\tau P} = S^{\uparrow^+ P}\) since \(\tau = \uparrow^+\). If \(P\) is substantial in \(S\), then, observing that \(S^{\tau P} \in \text{opt}^\circ_S BP\) by our initial assumption, we apply Lemma 65 to conclude that \(S^{\tau P} = S^{\uparrow P} = S^{\uparrow^+ P}\). 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 \(\uparrow^-\) (weak semi-trust) and \(\uparrow^-\) (semi-trust).

**Proposition 67.** Let \(\tau\) be a dynamic attitude.

- Suppose that \(\tau\) is positionally optimal for dual belief \(B^-\). Then \(\tau = \uparrow^-^+\).
- Suppose that \(\tau\) is positionally optimal for dual strong belief \(Sb^-\). Then \(\tau = \uparrow^-^+\).

As the proof is similar to the one given for \(\uparrow\), we do not provide details here.