A matter of trust: Dynamic attitudes in epistemic logic

Rodenhäuser, L.B.

Citation for published version (APA):
Chapter 2.  
Trust, Distrust, Semi-Trust

The purpose of this chapter is to identify various classes of dynamic attitudes that formalize reliability 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 $P$. 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. 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

1The 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.

2Stalnaker (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 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 $S$ iff best$S \subseteq P$; 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. 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.

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 an 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 $S$, and who (dynamically) accepts an input $P$ from a $\tau$-trusted source, will transform her epistemic state to an order $S^{\tau P}$ such that

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

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

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.,

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

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

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

**Proposition 13.** There exists no dynamic attitude $\tau$ satisfying success and sanity.
Towards a contradiction, let $\tau$ be a dynamic attitude satisfying success and sanity. Let $S$ be a non-empty plausibility order, and let $P$ be a proposition such that $P \cap S = \emptyset$. By success, $\text{best}^{\tau P} \subseteq P$. Since also $\text{best}^{\tau P} \subseteq S$, it follows that $\text{best}^{\tau P} \subseteq P \cap S$, thus $\text{best}^{\tau P} \subseteq \emptyset$, hence $S^{\tau P} = \emptyset$. However, since $S \neq \emptyset$, it follows from sanity that $S^{\tau P} \neq \emptyset$. We have arrived at a contradiction, hence $\tau$ 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 $S$ and proposition $P$:

$$\text{best}^{\tau P} \subseteq P \text{ and if } \neg (P \cap S) \neq \emptyset, \text{ then } S^{\tau P} \neq \emptyset.$$  

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:

$$\text{if } S \neq \emptyset, \text{ then } S^{\tau P} \neq \emptyset \text{ and if } \neg (P \cap S) \neq \emptyset, \text{ then } \text{best}^{\tau P} \subseteq P.$$  

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

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 $\tau$ is **positive** if the following holds, for any given order $S$ and proposition $P$:

$$\text{If } \neg (P \cap S) \neq \emptyset, \text{ then } S^{\tau P} \neq \emptyset \text{ and } \text{best}^{\tau P} \subseteq P.$$  

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 $\tau$, the upgrade $\tau P$ 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
2.1. Positive Attitudes

information of the agent before the upgrade. So, to reiterate, positive attitudes \( \tau \) 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 \( \tau \) be positive. For any order \( S \) and proposition \( P \), one of the following holds:

1. best\( S^{\tau P} \subseteq P \) and if \( P \cap S \neq \emptyset \), then \( S^{\tau P} \neq \emptyset \).
2. if \( S \neq \emptyset \), then \( S^{\tau P} \neq \emptyset \) and if \( P \cap S \neq \emptyset \), then best\( S^{\tau P} \subseteq P \).

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

If \( P \cap S = S \), then \( S^{\tau P} = S \).

**Proof.** Let \( \tau \) be positive and suppose that \( P \cap S = S \). If \( S = \emptyset \), then \( S^{\tau P} = \emptyset = S \), and the claim holds. If \( S \neq \emptyset \), it follows by definition of \( \tau \) that \( S^{\tau P} \neq \emptyset \). By strong informativity, \( S^{\tau P} \in \{ S, \emptyset \} \); so \( S^{\tau P} = S \).
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 \( \tau \) create the disjunction of belief and opposite knowledge, i.e., if \( \tau \) is positive, then for any order \( S \) and proposition \( P \): \( S \tau P \equiv BP \lor K \neg P \).

**Proof.** From left to right, let \( \tau \) be a positive attitude. If \( P \cap S = \emptyset \), then \( S \tau P = S \) or \( S \tau P = \emptyset \) by the informativity of dynamic attitudes. In either case, \( S \tau P \equiv K \neg P \) and the claim holds. Suppose now that \( P \cap S \neq \emptyset \). By definition of \( \tau \), this implies that \( S \tau P \neq \emptyset \) and \( \text{best} S \tau P \subseteq P \). Hence \( S \tau P \equiv BP \) 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 \( \tau \) is **strictly positive** iff the following are satisfied:

1. \( \text{best} S \tau P \subseteq P \).

2. If \( P \cap S \neq \emptyset \), then \( S \tau P \neq \emptyset \).

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 \( \tau \):

\[
\text{If } P \cap S = \emptyset, \text{ then } S \tau P = \emptyset.
\]

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
Strictly Positive Attitudes

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 unexecutable.

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

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

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

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

1. $\text{best} S^{* P} \subseteq P$. (success)

2. If $\text{best} S \cap P \neq \emptyset$, then $\text{best} S^{* P} = \text{best} S \cap P$. (expansion)

3. If $P \cap S \neq \emptyset$, then $S^{* P} \neq \emptyset$. (conditional sanity)

4. If $\text{best} S^{* P} \cap Q \neq \emptyset$, then $\text{best} S^{* (P \cap Q)} = \text{best} S^{* P} \cap Q$. (rational monotonicity)

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

---


\(5\) 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—, $\tau$ 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 $P \cap S \neq \emptyset$, then best $S^{\tau P} \subseteq P$.”

2.2.3. Examples. Infallible trust $!$ is an AGM operator; however, strong trust $\uparrow$ and minimal trust $\hat{\uparrow}$ 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 $\hat{\uparrow}$) 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 $\tau$ be a positive dynamic attitude. The stricture $\tau^+$ of $\tau$ is defined by means of

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

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 $\tau$, 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 $\tau$ is strictly positive iff $\tau$ is the stricture of some positive attitude.

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

that (2) if \( P \cap S \neq \emptyset \), then \( S^{\tau P} \neq \emptyset \). For (1), observe that if \( P \cap S \neq \emptyset \), then \( S^{\tau P} = S^\sigma P \), and since \( \sigma \) is positive, the claim holds. Further, if \( P \cap S = \emptyset \), then \( S^{\tau P} = \emptyset \) by definition of strictures, and again, the claim holds. For (2), observe that if \( P \cap S \neq \emptyset \), then \( S^{\tau P} = S^\sigma P \), and since \( \sigma \) is positive, the claim holds. So \( \tau \) 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 \( S \) be a plausibility order, and let \( P \) be a proposition.

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

2. If \( \text{best} S \cap P \neq \emptyset \), then \( \text{best} S^{\tau P} = \text{best} S \cap P \) by definition of \( \uparrow^+ \). So \( \uparrow^+ \) satisfies expansion.

3. If \( P \cap S \neq \emptyset \), then \( \text{best} S^{\tau P} P \neq \emptyset \), since in that case, \( S^{\tau P} = S \). So \( \uparrow^+ \) satisfies conditional sanity.

4. Finally, suppose that \( \text{best} S^{\tau P} \cap Q \neq \emptyset \). Thus, by definition of \( \uparrow^+ \), \( (\text{best}_S P) \cap Q \neq \emptyset \). Hence \( \text{best}_S (P \cap Q) = (\text{best}_S P) \cap Q \). By definition of \( \uparrow^+ \), \( \text{best} S^{\tau (P \cap Q)} = \text{best}_S (P \cap Q) \), but as we have just seen, \( \text{best}_S (P \cap Q) = (\text{best}_S P) \cap Q \), so \( \text{best} S^{\tau (P \cap Q)} = \text{best} S^{\tau P} \cap Q \). 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 \( i \), called weak infallible trust, by means of

\[
S^{i P} := \begin{cases} 
S^{\tau P} & \text{if } P \cap S \neq \emptyset, \\
S & \text{if } P \cap S = \emptyset.
\end{cases}
\]

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 viewed 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 distrust them.

A Smullyan liar claims: I don’t live on this island.

As anyone familiar with Smullyan’s logic puzzles knows, 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 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 $\tau$ is negative if (for any $S$ and $P$)

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

On the other hand, an attitude $\tau$ is strictly negative if the following holds:

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

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

\[\text{Cf., for example, Smullyan (1978).}\]
2.3. Negative Attitudes

PROPOSITION 21.

1. For any negative attitude \( \tau \): If \( \neg P \cap S = S \), then \( S^{\tau P} = S \).

2. For any strictly negative attitude \( \tau \): If \( \neg P \cap S = \emptyset \), then \( S^{\tau P} = \emptyset \).

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

PROPOSITION 22.

1. Negative attitudes \( \tau \) create the disjunction of disbelief and knowledge, i.e., if \( \tau \) is negative, then for any plausibility order \( S \) and proposition \( P \): \( S^{\tau P} = B_{\neg P} \lor KP \).

2. Strictly negative attitudes \( \tau \) create disbelief, i.e., if \( \tau \) is strictly negative, then for any plausibility order \( S \) and proposition \( P \): \( S^{\tau P} = B_{\neg P} \).

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 \( \tau^{-} \) of an attitude \( \tau \) is given by \( \tau^{-} := \tau(\neg P) \).

PROPOSITION 23. An attitude \( \tau \) is (strictly) positive iff its opposite \( \tau^{-} \) is (strictly) negative.

PROOF. We show the claim for strictly positive attitudes; the claim for positive attitudes is shown analogously. Suppose that \( \tau \) is strictly positive, and consider a plausibility order \( S \) and a proposition \( P \). Since \( \tau \) is strictly positive, \( \text{best} S(\neg P) \subseteq \neg P \). Furthermore, if \( \neg P \cap S \neq \emptyset \), then \( S^{\tau(\neg P)} \neq \emptyset \). Recalling that \( \tau^{-} = \tau(\neg P) \) for any \( P \), it follows that \( \tau^{-} \) 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 \( !^{-} \), strong distrust is given by \( \uparrow^{-} \), strict strong distrust is given by \( \uparrow^{-} := (\downarrow^{+})^{-} \), minimal distrust is given by \( \uparrow^{-} \), and strict minimal distrust is given by \( \uparrow^{-} := (\downarrow^{+})^{-} \). Of these, \( \downarrow^{-} \) and \( \uparrow^{-} \) are negative, while \( !^{-} \), \( \uparrow^{-} \) 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 $\tau$ is *semi-positive* iff:

If $P \cap S \neq \emptyset$, then $\text{best}_{S^p} \cap P \neq \emptyset$.

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 $P \cap S$ is non-empty, then
also \( \text{best}_{S\cap P} \cap P \) 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 \( \tau \)-source will result in suspension of disbelief, i.e., applying the upgrade \( \tau P \) will result in an order satisfying \( B^\sim P \), the dual of belief in \( P \) (cf. §\ref{sec:1.2.10}) given by

\[
S \equiv BP \iff \text{best} S \cap P \neq \emptyset.
\]

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\[\text{6}\] 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 \( \tau \) is semi-negative iff: if \( \neg P \cap S \neq \varnothing \), then \( \text{best}_S^{\tau P} \cap \neg P \neq \varnothing \).

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

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

PROOF. We prove the first item as an example. Let \( S \) be a plausibility order, and \( P \) a proposition. Suppose that \( \neg P \cap S \neq \varnothing \). By definition of \( \tau \), \( S^{\tau P} = S^{\tau \neg P} \), with \( \tau \neg \) semi-positive. Since \( \neg P \cap S \neq \varnothing \) and \( \tau \neg \) is semi-positive, we conclude that \( \text{best}_S^{\tau \neg P} \cap \neg P \neq \varnothing \). But since \( S^{\tau \neg P} = S^{\tau P} \), 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 \( B^\neg \), while strictly semi-positive attitudes create the dual of belief \( B^\neg \):

PROPOSITION 25.

1. Semi-positive attitudes \( \tau \) create the disjunction of opposite knowledge and dual belief, i.e., if \( \tau \) is semi-positive, then for any plausibility order \( S \) and proposition \( P \): \( S^{\tau P} \models K^\neg P \lor B^\neg P \).

2. Positive attitudes \( \tau \) create dual belief, i.e., if \( \tau \) is semi-positive, then for any plausibility order \( S \) and proposition \( P \): \( S^{\tau P} \models B^\neg P \).

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 \( B^\neg P \) and \( B^\neg \neg P \) 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 \( \tau \) is an AGM contraction operator iff for any order \( S \) and proposition \( P \):

- \( \text{best} S \subseteq \text{best} S^{\tau P} \). (inclusion)
- If \( \text{best} S \cap \neg P \neq \emptyset \), then \( \text{best} S^{\tau P} = \text{best} S \). (vacuity)
- If \( S \cap \neg P \neq \emptyset \), then \( \text{best} S^{\tau P} \cap \neg P \neq \emptyset \). (success of contraction)
- If \( \text{best} S \subseteq P \), then \( \text{best} S^{\tau P} \cap \text{best} S \cap P \subseteq \text{best} S \). (recovery)
- \( \text{best} S^{\tau (P \cap Q)} \subseteq \text{best} S^{\tau P} \cup \text{best} S^{\tau Q} \). (conjunctive overlap)
- If \( \text{best} S^{\tau (P \cap Q)} \cap \neg P \neq \emptyset \), then \( \text{best} S^{\tau P} \subseteq \text{best} S^{\tau (P \cap Q)} \). (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 \( \hat{\sim} \neg \) (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 \( \rightarrow \) on dynamic attitudes by means of

\[ S \rightarrow S' \text{ iff } S' \subseteq S. \]

Equivalently, \( S \rightarrow S' \) iff there exists an upgrade \( U \) such that \( S^U = S' \). If \( S \rightarrow S' \), then we say that \( S' \) is in the dynamic scope of \( S \).

Essentially, \( S' \) is in the dynamic scope of \( S \) if \( S' \) is reachable from \( S \) “in theory.” But is \( S' \) also reachable from \( S \) “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 attitudes towards the sources are collected in the set $\Delta$. An interesting question to consider is whether any order $S'$ that is in the dynamic scope of some given order $S$ ("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 $\tau_0P_0, \ldots, \tau_nP_n$, where each $\tau_k$ is an element of $\Delta$ ("reachable in practice"). If this is the case in general, then the set $\Delta$ 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 $\Delta$, we write $S \rightarrow_\Delta S'$ iff there exists an attitude $\tau \in \Delta$ and a proposition $P \subseteq W$ such that $S^{\tau P} = S'$. Notice that whenever $S \rightarrow_\Delta S'$ for some $\Delta$, then also $S \rightarrow S'$.

We say that a set of dynamic attitudes $\Delta$ is (dynamically) complete if for any plausibility orders $S$ and $S'$ such that $S \rightarrow S'$ there exists a sequence of plausibility orders $S_0, \ldots, S_n$ ($n \geq 0$) such that $S \rightarrow_\Delta S_0 \rightarrow_\Delta \cdots \rightarrow_\Delta S_n \rightarrow_\Delta S'$. Intuitively, this boils down to the picture painted right above: suppose an agent is in the epistemic state given by $S$, and suppose that the set $\Delta$ collects the agent’s sources of information. Take any order $S'$ such that $S \rightarrow S'$. 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 $S'$. If this is the case for arbitrary pairs $S$ and $S'$ such that $S \rightarrow S'$, then the set of dynamic attitudes $\Delta$ 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 $\Delta$ be a set of dynamic attitudes. $\Delta$ is monotonic iff for any order $S$, proposition $P$ and $\tau \in \Delta$: if $S^{\tau P} \neq S$, then it is not the case that $S^{\tau P} \rightarrow_{\Delta} S$.

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 $\Delta$ be a set of dynamic attitudes. If $\Delta$ is monotonic, then $\Delta$ is incomplete.

**Proof.** We show the contrapositive: if $\Delta$ is complete, then $\Delta$ is not monotonic. Suppose that $\Delta$ is complete. Let $S, S'$ be plausibility orders such that $S \neq S'$ and $S = S'$. By the fact that $\Delta$ is complete, we have $S \rightarrow_\Delta S'$, so
there exists a sequence of upgrades \( \tau_1 P_1 \ldots \tau_n P_n \) of minimal length such that 

\[ (...(S^{\tau_1 P_1})\ldots)^{\tau_n P_n} = S' \]

with \( \tau_k \in \Delta \), and \( P_k \in W \) for \( 1 \leq k \leq n \). Note that \( n \geq 1 \) since \( S \neq S' \). Since our sequence is of minimal length, 

\[ (...(S^{\tau_1 P_1})\ldots)^{\tau_{n-1} P_{n-1}} \neq S'. \]

Since \( \Delta \) is complete, \( S' \not\rightarrow_{\Delta} (...(S^{\tau_1 P_1})\ldots)^{\tau_{n-1} P_{n-1}} \). So \( \Delta \) 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

\[ S = \{(w, v), (w, v), (w, w), (v, v)\} \]

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

**Proof.** We assume the notation from the statement of the lemma. Suppose first that \( P \cap S = \emptyset \). By informativity of dynamic attitudes, \( S^{\tau P} \in \{\emptyset, S\} \). Since \( S \neq \emptyset \), it follows that \( S^{\tau P} = S \), and our claim holds. Suppose, second, that \( P \cap S \neq \emptyset \). We distinguish two sub-cases. First, suppose that \( \{w, v\} \subseteq P \). Then \( S^{\tau P} = S \) by Proposition [15] and our claim holds. Second, suppose that \( \{w, v\} \notin P \). This implies that either (1) \( w \in P, v \notin P \) or (2) \( w \notin P, v \in P \). Assuming (1), it follows that \( (w, v) \in S^{\tau P}, (v, w) \notin S^{\tau P} \) since \( \tau \) is positive. But then, \( S^{\tau P} = S \). Assuming (2), it follows that \( (v, w) \in S^{\tau P}, (w, v) \notin S^{\tau P} \), again, since \( \tau \) is positive. But then \( S^{\tau P} = S' \). 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 \( \Delta \) be a set of positive attitudes. Consider the order

\[ S = \{(w, v), (w, v), (w, w), (v, v)\} \]

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

\[ (...(S^{\tau_1 P_1})\ldots)^{\tau_n P_n} \in \{S, S'\} \]

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

\[ S'' = \{(w, v), (v, w), (w, w), (v, v)\} \]

Thus it is not the case that \( S \not\rightarrow_{\Delta} S'' \). However, \( S \not\rightarrow S'' \). So \( \Delta \) is incomplete.
Chapter 2. Trust, Distrust, Semi-Trust

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 \( \Delta = \{!, \uparrow, \uparrow^-\} \) is dynamically complete. Since \( \uparrow, \uparrow^- \) and ! are semi-positive, this entails the original claim. Let \( S, S' \) be plausibility orders such that \( S \rightarrow S' \). We give a sketch how to obtain \( S' \) from \( S \) using !, \( \uparrow \) and \( \uparrow^- \): first, upgrade \( S \) with \(!S'\); second, make all worlds in \( S' \) equiplausible using a sequence of upgrades given by \( \uparrow^- \) 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 \( S' \) 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 \( S' \), then the set of worlds that are second-to-least plausible in \( S' \), and so on, and in this way we are guaranteed to finally arrive at \( S' \) itself). As a result, \( S \rightarrow_\Delta S' \). Hence \( \Delta = \{\uparrow, \uparrow^-, !\} \) is dynamically complete.

The construction of \( S' \) from \( S \) crucially involves the use of \( \uparrow^- \) 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 \( B^- \); and strict semi-positive
attitudes (which are just the strictures of semi-positive attitudes) create dual belief $B^-$.

An analogous style of analysis may now be applied to other propositional attitudes, in particular, to the two main other ones we have seen: irrevocable 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 $\tau$ (the last row in Table 2), the propositional input $P$ is “accepted” only insofar as applying the upgrade $\tau P$ to an order $S$ yields an order $S^{\tau P}$ which still contains $P$-worlds, provided $S$ 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”\footnote{Upwards refinement was introduced under the name “refinement” by Papini (2001).} given a plausibility order $S$, we scan $S$ 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 illustrates what is going on in
Proceeding to a formal definition, the *upwards refinement* $\up$ is the dynamic attitude defined by setting $S^{\up P} = S$, and stipulating, for any $w, v \in S$, that $w \leq_{S^{\up P}} v$ iff

- $(w \in P \iff v \in P)$ and $w \leq_S v$, or
- $w \in P, v \notin P$ and $w \leq_S v$, or
- $w \notin P, v \in P$ and $w < S v$.

Rephrasing the above: this simply means that for each $w \in S$, we split the set $\{v \in S \mid v \approx_S w\}$ 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 $\up$ is the propositional attitude *refinedness*, denoted by $R$ and given by

$$S \models R P \iff \forall w, v \in S: \text{ if } w \in P \text{ and } v \notin P, \text{ then } w <_S v \text{ or } v <_S w.$$ 

In terms of upwards refinement $\up$ and minimal trust $\uparrow$, we will define a dynamic attitude which we call *moderate trust*, denote with $\ll$, and which will turn out to be intermediate in strength between minimal trust $\uparrow$ and strong trust $\uparrow \ll$. To do so, we first introduce another operation on dynamic attitudes, called the “composite”. A pair of dynamic attitudes $(\sigma, \tau)$ 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 $\sigma$ and $\tau$ is given by

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

Now: *moderate trust* $\ll$ is the composite of $\up$ and minimal trust $\uparrow$, that is:

$$\ll := \up \cdot \uparrow$$

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., $\ll P = \up P \cdot \uparrow P$, for any $P \subseteq W$. Observe also that $\up \cdot \uparrow = \uparrow \cdot \up$; the order does not matter.

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

---

8Moderate trust has first been considered under the name of “restrained revision” by [Booth and Meyer](2006).
2.5. Qualitative Degrees of Trust and Semi-Trust

After performing an additional upgrade with any $Q$ such that $\text{best}_{S^P} Q \cap P \neq \emptyset$. This is not the case for minimal trust $\uparrow!$. In terms of the subsumption order, we have $\uparrow\uparrow < \uparrow$, as can be seen by noting that the fixed point of $\uparrow\uparrow$, 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 $S$ amounts to making the best $P$-worlds as good as the worst non-$P$-worlds, while otherwise preserving the pairs given by $S$. If there are no $P$-worlds, or no non-$P$-worlds, nothing changes, i.e., in that case, $S \uparrow\sim P = S$.

2.5.3. BARE SEMI-TRUST. An even weaker dynamic attitude is given by bare semi-trust $\uparrow\sim$, defined as follows:

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

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

$$
S \models K^\sim P \text{ iff } S \cap P \neq \emptyset.
$$

We can interpret $\uparrow\sim$ 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).

---

9An analogous notion was introduced by Veltman (1996) to give a semantics for the epistemic modal might. We return to this in Chapter 6.
Chapter 2. Trust, Distrust, Semi-Trust

<table>
<thead>
<tr>
<th>Semantic Condition</th>
<th>Characterization</th>
<th>Typical Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^P \neq \emptyset$ and $\text{best} \ S^P \subseteq P$.</td>
<td>–</td>
</tr>
<tr>
<td><strong>Strictly Positive</strong></td>
<td>$\text{best} \ S^P \subseteq P$ and if $P \cap S \neq \emptyset$, then $S^P \neq \emptyset$.</td>
<td>Strictures of positive attitudes</td>
</tr>
<tr>
<td><strong>Negative</strong></td>
<td>If $S \cap \neg P \neq \emptyset$, then $S^P \neq \emptyset$ and $\text{best} \ S^P \subseteq \neg P$.</td>
<td>Opposites of positive attitudes</td>
</tr>
<tr>
<td><strong>Strictly negative</strong></td>
<td>$\text{best} \ S^P \subseteq \neg P$ and if $S \cap \neg P \neq \emptyset$, then $S^P \neq \emptyset$.</td>
<td>Opposites of strictly positive attitudes</td>
</tr>
<tr>
<td><strong>Semi-positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^P \neq \emptyset$ and $\text{best} \ S^P \cap P \neq \emptyset$.</td>
<td>–</td>
</tr>
<tr>
<td><strong>Semi-negative</strong></td>
<td>If $\neg P \cap S \neq \emptyset$, then $S^P \neq \emptyset$ and $\text{best} \ S^P \cap \neg P \neq \emptyset$.</td>
<td>Opposites of semi-positive attitudes</td>
</tr>
</tbody>
</table>

**Table 1.** Classes of dynamic attitudes related to simple belief.
2.5. Qualitative Degrees of Trust and Semi-Trust

<table>
<thead>
<tr>
<th>Definition of class</th>
<th>Strictures create</th>
<th>Typical example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Irrevocably positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \neq \emptyset$ and $S^\tau P \subseteq P$</td>
<td>irrevocable knowledge $K$</td>
</tr>
<tr>
<td><strong>2. Strongly positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \neq \emptyset$ and for all $w, v \in S^\tau P$: if $w \in P$, $v \notin P$, then $w &lt;_{S^\tau P} v.$</td>
<td>strong belief $Sb$</td>
</tr>
<tr>
<td><strong>3. Moderately positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \neq \emptyset$, best $S^\tau P \subseteq P$ and for all $w, v \in S^\tau P$: if $w \in P, v \notin P$ and $w \leq_s v$, then $w &lt;_{S^\tau P} v.$</td>
<td>refined belief $Rb$</td>
</tr>
<tr>
<td><strong>4. Positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \neq \emptyset$ and best $S^\tau P \subseteq P$</td>
<td>simple belief $B$</td>
</tr>
<tr>
<td><strong>5. Semi-positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \neq \emptyset$ and best $S^\tau P \cap P \neq \emptyset$</td>
<td>dual belief $B^\sim$</td>
</tr>
<tr>
<td><strong>6. Weakly semi-positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then (1) $S^\tau P \neq \emptyset$ and (2) if $S^\tau P \cap \neg \neq \emptyset$, then exist $w, v \in S^\tau P$: $w \in P, v \notin P$: $w \leq_{\tau P} v$</td>
<td>dual strong belief $Sb^\sim$</td>
</tr>
<tr>
<td><strong>7. Barely positive</strong></td>
<td>If $P \cap S \neq \emptyset$, then $S^\tau P \cap P \neq \emptyset$</td>
<td>dual knowledge $K^\sim$</td>
</tr>
</tbody>
</table>

*Table 2. Seven Qualitative Degrees of Trust and Semi-Trust.*
<table>
<thead>
<tr>
<th>Reminder for Definition</th>
<th>Fixed point of stricture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Infallible trust</strong> !</td>
<td>“Delete all non-P-worlds”</td>
</tr>
<tr>
<td><strong>2. Strong trust</strong> ( \uparrow )</td>
<td>“Make all P-worlds strictly more plausible than all non-P-worlds”</td>
</tr>
<tr>
<td><strong>3. Moderate Trust</strong> ( \uparrow \uparrow )</td>
<td>“Compose minimal trust with refinement”</td>
</tr>
<tr>
<td><strong>4. Minimal trust</strong> ( \uparrow )</td>
<td>“Make the best P-worlds the best worlds overall”</td>
</tr>
<tr>
<td><strong>5. Semi-trust</strong> ( \uparrow^- )</td>
<td>“Add the best P-worlds to the best worlds overall”</td>
</tr>
<tr>
<td><strong>6. Weak semi-trust</strong> ( \uparrow^- )</td>
<td>“Make the best P-worlds at least as good as the worst non-P-worlds”</td>
</tr>
<tr>
<td><strong>7. Bare semi-trust</strong> ( \uparrow^- )</td>
<td>“Test whether there are any P-worlds”</td>
</tr>
</tbody>
</table>

**Table 3.** Seven Typical Examples of Dynamic Attitudes.
2.5. Qualitative Degrees of Trust and Semi-Trust

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 $\up^\sim$: applying an upgrade $\up^\sim P$ 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. $\uparrow = K$ (the fixed point of infallible trust is irrevocable knowledge).
2. $\uparrow^\uparrow = B$ (the fixed point of strict minimal trust is belief).
3. $\uparrow^\uparrow^\uparrow = S_b$ (the fixed point of strict strong trust is strong belief).
4. $\uparrow^\uparrow^\uparrow = R_b$ (the fixed point of strict moderate trust is refined belief).
5. $\uparrow^\uparrow^\uparrow = B^\sim$ (the fixed point of strict semi-trust is the dual of belief).
6. $\uparrow^\uparrow^\uparrow = S_b^\sim$ (the fixed point of strict weak semi-trust is the dual of strong belief).
7. $\cong = K^-$ (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:

PROPOSITION 31.

1. $\emptyset < ! < \uparrow < \uparrow \uparrow < \uparrow^+ < \uparrow^\sim < !^- < \text{id}.$

2. $\emptyset < ! < \uparrow^+ < \uparrow^\sim < \uparrow^\sim < !^- < \text{id}.$

PROOF. We give just one example. To show that $\uparrow < \uparrow$, we need to verify that for any plausibility order $S$ and proposition $P$, $(S\uparrow^P)^\uparrow P = S\uparrow^P$. Assuming that $P \cap S = \emptyset$, it follows that $S\uparrow^P = S = S\uparrow^P$, so our claim holds. Suppose, then, that $P \cap S = \emptyset$. Then for any $w, v \in S$ such that $w \in P, v \notin P$, we have that $w <_{SP} v$. In other words: there are no ties between $P$-worlds and non-$P$-worlds in $S\uparrow^P$. Hence $(S\uparrow^P)^\uparrow P = S\uparrow^P$. 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 $S$ and a proposition $P$ such that $P \cap S = \emptyset$. In such a situation, $S\uparrow^P = S$, however, $S^\uparrow^PP = \emptyset$, hence $(S\uparrow^P)^\uparrow^P \neq S\uparrow^P$, so $\not\exists w \in S^\uparrow^P v$. 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 $\tau^+$ rather than $\tau$: otherwise, our argument does not go through).

PROPOSITION 32.

1. For all extremely positive attitudes $\tau$: $\tau^+ \leq !$.

2. For all strongly positive attitudes $\tau$: $\tau \leq \uparrow$.

3. For all moderately positive attitudes $\tau$: $\tau \leq \uparrow \uparrow$.

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

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

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 $\tau$: $\tau \leq \uparrow^\sim$.

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

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

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

2. Let $\tau$ be a strongly positive attitude. Let $S$ be a plausibility order and $P$ a proposition. Since $\tau$ is strongly positive, $S^{\tau P} = (K^\sim \lor Sb) P$. Since the fixed point of $\uparrow$ is $K^\sim \lor Sb$, $(S^{\tau P})^{\uparrow} = S^{\tau P}$. So $\tau \leq \uparrow$.

3. Let $\tau$ be a moderately positive attitude. Let $S$ be a plausibility order and $P$ a proposition. Since $\tau$ is moderately positive, $S^{\tau P} = K^\sim P \lor RbP$. Since the fixed point of $/uni21c8$ is $K^\sim \lor Rb$, $(S^{\tau P})^{/uni21c8} = S^{\tau P}$. So $\tau \leq /uni21c8$.

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 irrevocable knowledge $?K$, which, by definition (cf. §1.7.4), is given by

$$S^{?KP} := \begin{cases} S & S \models KP, \\ \emptyset & \text{otherwise}. \end{cases}$$

The fixed point of $?K$ is obviously irrevocable knowledge $K$, which is also the fixed point of infallible trust $!$: $?K = !$. Hence, by Theorem 14, $?K = !$ (i.e., $?K$ and $!$ mutually subsume each other, cf. §1.8.2). Thus, by Proposition 32 above, for any extremely positive attitude $\tau$: $\tau \leq ?K$.

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 attitude corresponding to irrevocable knowledge $K$? What makes the connection between infallible trust $!$ and irrevocable knowledge $K$ unique? Analogous questions may be asked for $\uparrow$, $\uparrow^\sim$ 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 \( \sigma \) and \( \tau \) contingent on whether the current epistemic state supports a particular propositional attitude \( A \). 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 \( u_{APv} \) by means of

\[
S_{u_{APv}} := \begin{cases} 
S^u & \text{if } S \models AP \\
S^v & \text{if } S \not\models AP
\end{cases}
\]

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

— A pair of attitudes \((\sigma, \tau)\) is \(A\)-mixable if the family of upgrades \(\{\sigma P\tau P\}_P\subseteq W\) is an attitude\(^{12}\).

— Given dynamic attitudes \(\sigma\) and \(\tau\), and a propositional attitude \(A\), we define the mixture \(\sigma_A \tau\) of \(\sigma\) and \(\tau\) over \(A\) by means of

\[
S(\sigma_A \tau)_P := \begin{cases} 
S^\sigma P_A S^\tau P \quad & (\sigma, \tau) \text{ is } A\text{-mixable} \\
\emptyset & \text{otherwise} 
\end{cases}
\]

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

\[
S \models DP \text{ iff } S \models BQ.
\]

So the proposition \(P\) that \(D\) takes an argument is actually ignored: all that matters for \(DP\) to satisfied in a plausibility order \(S\) is whether the agent believes he is drunk in \(S\). 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 \(\uparrow_{\neg D}\text{id}\), the mixture of minimal trust \(\uparrow\) and neutrality \(\text{id}\) 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 \((\sigma, \tau)\) is \(A\)-mixable iff for any order \(S\), 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.
— $S \models AP$ and $S^{\sigma P} = (AP \lor \tau P)$, or
— $S \models \neg AP$ and $S^{\tau P} = (\neg AP \lor \sigma P)$.

**Proof.** From left to right, suppose that $(\sigma, \tau)$ is $A$-mixable. Let $S$ be a plausibility order, and let $P$ be a proposition. We first assume that $S \models AP$. It follows that $S^{(\sigma \lor \tau) P} = S^{\sigma P}$. If $S^{\sigma P} \not\models AP$, our claim holds, so suppose that $S^{\sigma P} /\models AP$. Towards a contradiction, suppose that also $S^{\tau P} /\models AP$. Then $(S^{\sigma P})^{(\sigma \lor \tau) P} = (S^{\sigma P})^{\tau P} /\neq S^{\sigma P}$. Hence $(S^{\sigma \lor \tau P})^{(\sigma \lor \tau) P} /\neq S^{\sigma \lor \tau P}$. Thus the family of upgrades $\{\sigma \lor \tau P\}_{P \in W}$ is not a dynamic attitude. So $(\sigma, \tau)$ is not $A$-mixable. But this contradicts the initial assumption. We conclude that $S \models \tau P$, so our claim holds. Second, we assume that $S \models \neg AP$, 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 $\{\sigma P \lor \tau P\}_{P \in W}$ is an attitude. We show that $\sigma P^{\lor} \lor \tau P$ is idempotent. Pick an order $S$ and a proposition $P$. Suppose that $S \models AP$. Then $S^{\sigma P \lor \tau P} = S^{\sigma P}$. If $S^{\sigma P} /\models AP$, we are done, since $\sigma P$ is idempotent. If $S^{\sigma P} /\models \neg AP$, by the assumption, $S \models \tau P$, hence $(S^{\sigma P})^{(\sigma \lor \tau) P} = (S^{\sigma P})^{\tau P} = S^{\sigma P}$, which shows the claim. Under the assumption that $S \not\models AP$, 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 $\Gamma$. Given two dynamic attitudes $\sigma$ and $\tau$, we would then like to define a dynamic attitude $\nu$ such that

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

This means that if $P$ is “on topic” we use $\sigma$, and if $P$ is “off topic” we use $\tau$.

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

$$S \models \check{\Gamma} P \iff P \cap S \in \Gamma.$$  

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

2.6. Mixtures of Dynamic Attitudes

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 $\Gamma$ collects all "mathematical" propositions. Then we may capture this dynamic attitude by means of the mixture $\uparrow \Gamma \text{id}$ (the mixture of strong trust $\uparrow$ and neutrality $\text{id}$ over $\Gamma$). 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 $\tau$ given by

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

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

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

$$S^{\nu P} := \begin{cases} S^{\sigma P} & S^{\sigma P} \not\models \emptyset, \\ S^{\tau P} & \text{otherwise.} \end{cases}$$

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

$$S \models \checkmark_{\sigma} P \text{ iff } S^{\sigma P} \not\models \emptyset;$$

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

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

and we notice that, indeed, $\nu = \sigma \ll \tau$. 
3. Recall (§1.7.4) that for any propositional attitude \( A \), the test for \( A \), denoted by \( ?A \), is the dynamic attitude given by

\[
S^{?AP} := \begin{cases} S & S \vDash AP, \\ \emptyset & \text{otherwise.} \end{cases}
\]

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

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 \( \tau \) which, given another dynamic attitude \( \sigma \) and a propositional attitude \( A \), is defined by

\[
S^{\tau P} := \begin{cases} S^{\tau P} & S \vDash AP \\ \emptyset & \text{otherwise} \end{cases}
\]

This may be captured by a mixture, which we call the restriction of \( \tau \) to \( A \), denoted \( \tau_A \), and defined by \( \tau_A := \tau \emptyset \). 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 \( \tau \), the stricture \( \tau^+ \) of \( \tau \) by means of

\[
S^{\tau^+ P} := \begin{cases} S^{\tau P} & S \cap P \neq \emptyset \\ \emptyset & \text{otherwise,} \end{cases}
\]

The stricture of \( \tau \) is just the restriction of \( \tau \) to \( K^- \): \( \tau^+ = \tau_{K^-} \). 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
2.7. Mutual Assessments of Reliability

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. 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 $A$ (the set of agents). A multi-agent plausibility order (over $A$) is a family of agent-indexed preorders (i.e., reflexive, transitive relations)

$$\{S_a := (S, \leq^a_S)\}_{a \in A},$$

containing one preorder $S_a$ on $S$ for each agent $a \in A$.

A multi-agent plausibility order $\{S_a := (S, \leq^a_S)\}_{a \in A}$ differs from a single-agent plausibility order $S$ (cf. §1.1.1) in two respects: first, and most obviously, a multi-agent plausibility order $\{S_a := (S, \leq^a_S)\}_{a \in A}$ gives us a collection of preorders $S_a$, one for each agent $a \in A$. 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 $(\{S_a\}_{a \in A})$, let, for each agent $a$, the relation $\sim^a_S$ be given by:

$$w \sim^a_S v \text{ iff } w \leq^a_S v \text{ or } v \leq^a_S w.$$ 

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

---

14Cf. 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).

15Notice 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 \(v\).

Now for each world \(w \in S\), let \(a_S(w) := \{v \in S | w \sim^a_S v\}\). We call \(a_S(w)\) the information cell of \(a\) at \(w\) in \(\{S_a\}_{a \in A}\). It represents the agent’s hard information at \(w\) (in \(\{S_a\}_{a \in A}\)). Put differently, \(a_S(w)\) 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 \(\sim^a_S\) is, clearly, an equivalence, it follows that whenever \(w \sim^a_S v\), then \(a_S(w) = a_S(v)\).

2.7.3. Local States. Each information cell \(a_S(w)\) induces an ordering given by

\[ S_{a(w)} := S_a \cap (a_S(w), a_S(w) \times a_S(w)) \]

We call \(S_{a(w)}\) the local state of agent \(a\) at \(w\) in \(\{S_a\}_{a \in A}\).

The domain of the local state \(S_{a(w)}\) consists of the information cell \(a_S(w)\) of agent \(a\) at \(w\); and the relational pairs in the local state \(S_{a(w)}\) are given by those \((w, v) \in S_a\) such that \(w, v \in a_S(w)\).

Besides the “appearance” of the world \(w\) to agent \(a\) (given by \(a_S(w)\)), 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 \(a_S(w)\) in terms of \(\sim^a_S\), which is in turn defined in terms of \(\leq^a_S\), the order \(S_{a(w)}\) is reflexive, transitive and connected, so \(S_{a(w)}\) is, in fact, a (single-agent) plausibility order in the sense of §1.1.

Notice that this plausibility order, the local state \(S_{a(w)}\), presents the “appearance” of \(w\) to agent \(a\) viewed in isolation from other agents. If other agents are taken into account, then \(S_{a(w)}\) 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 \(S_a\) that live in a multi-agent plausibility order \(\{S_a\}_{a \in A}\) 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.

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

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

- \(a_S(x) = \{x\}, b_S(x) = \{x, w, y\}\),
- \(a_S(y) = \{w, y\}, b_S(y) = \{x, w, y\}\),
- \(a_S(w) = \{w, y\}, b_S(w) = \{x, y, w\}\).

So while at any of the three worlds \(w, x, y\), 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 \(w, y\), 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 \(S_{a(y)}\) of agent \(a\) induced by the cell \(a_S(x)\) is given by the single-agent plausibility order

\[S_{a(y)} = (a_S(y), \{(y, w), (w, y), (w, w), (y, y)\})\]

while \(S_{b(y)}\) (the local state of agent \(b\) induced by the cell \(a_S(y)\)) is given by the single-agent plausibility order

\[S_{b(y)} = (a_S(y), \{(w, x), (w, y), (x, y), (y, x), (w, w), (x, x), (y, y)\})\]

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
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 $\tau$ 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 9 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 $\uparrow \Delta \Gamma id$ (the mixture of strong trust and neutrality over $\Delta \Gamma$, with $\Gamma$ 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 $a, b \in A$ such that $a \neq b$: there exists a dynamic attitude $\tau$ such that $(a, \tau, b) \in T$ (existence).

2. For any two agents $a, b \in A$ 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
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 \( \{S_a\}_{a \in A} \) as defined in §2.7.1 above. We would like to formally encode some additional information, answering the question how each agent \( b \in A \) assesses the reliability of each other agent \( c \in A \) (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 \( \{ S_a \}_{a \in A} \) a multi-agent plausibility order. A trust labeling over \( \{ S_a \}_{a \in A} \) is a function \( T \) assigning to each possible world in \( W \) a trust graph \( T_w \), satisfying for each agent \( b, c \in A \):

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

So a trust labeling \( T \) gives us a trust graph \( T_w \) for each possible world \( w \in S \), with the additional requirement stated 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

\[(\{ S_a \}_{a \in A}, T)\]

where \( \{ S_a \}_{a \in A} \) is a multi-agent plausibility order, and \( T \) is a trust labeling over \( \{ S_a \}_{a \in A} \).

2.7.7. TRUST-PLAUSIBILITY TRANSFORMER. A trust-plausibility transformer \( [c] \) is a function

\[ C \xrightarrow{[c]} C[c] \]

assigning a trust-plausibility order

\[ C[c] := (\{ S_a[c] := (S[c], \leq_a S_a[c]) \}_{a \in A}, T[c]) \]

to each given trust-plausibility order \( C = (\{ S_a \}_{a \in A}, T) \), in such a way that \( S[c] \subseteq S \) 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 \( C \) into a new joint information state, given by the trust-plausibility model \( C[c] \).

Given a world \( w \in S \), the trust-plausibility transformer \( [c] \) is executable in \( C \) at \( w \) iff \( w \in S[c] \).

Notice that to uniquely determine a trust-plausibility transformer \( [c] \), it is enough to specify, for each trust-plausibility order \( (\{ S_a \}_{a \in A}, T) \), what \( [c] \) does to the underlying multi-agent plausibility order \( \{ S_a \}_{a \in A} \), 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 \( S_a[c] \), for each agent \( a \) and for each given trust-plausibility order \( (\{ S_a \}_{a \in A}, T) \), defines a communication act \( [c] \).
2.7. Mutual Assessments of Reliability

Keeping the trust labeling fixed on application of a trust-plausibility transformer encodes the assumption that agents do not change their mutual assessments of reliability 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 \(b \in A\), 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 \(C = (\{S_a\}_{a \in A}, T)\), agent \(b \in A\), and proposition \(P \subseteq W\), we would like to specify a multi-agent plausibility order \(\{S_a[b: P]\}\). 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 \(a \in A\) and each possible world \(w \in S\), the local state \(S_a(w)\) 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 \(C = (\{S_a\}_{a \in A}, [\cdot], L)\) and a world \(w \in S\), we make the notation
\[
\tau^w_{a \rightarrow b} := T_w(a, b)
\]

Start now with the new domain \(S[b: P]\). 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:

\[
S[b; P] := \{ w \in S \mid \forall a \neq b \in A : w \in (S_{b(w)})^{\tau_{a\rightarrow b}P} \}.
\]

We write \( S[b; P] \) for the natural product order on \( S[b; P] \), that is,

\[
S[b; P] := (S[b; P], S[b; P] \times S[b; P]).
\]

To obtain the orderings on the new domain \( S[b; P] \) 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 \( S[b; P] \). So for any \( a \neq b \), put:

\[
S_a[b; P] := \left( \bigcup_{w \in S} (S_{b(w)})^{\tau_{a\rightarrow b}P} \right) \cap S[b; P]
\]

For agent \( b \), we simply put

\[
S_b[b; P] := S_b \cap S[b; P],
\]

i.e., we restrict the old order \( S_b \) to the new domain \( S[b; P] \).

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

\[
C[b; P] := (\{S_a[b; P]\}_{a \in A}, T),
\]

given by:

\[
S[b; P] := \{ w \in S \mid \forall a \neq b \in A : w \in (S_{b(w)})^{\tau_{a\rightarrow b}P} \},
\]

where we assume that \( T_w(a, b) = \tau_{a\rightarrow b} \), and

\[
\forall a \neq b : S_a[b; P] := \left( \bigcup_{w \in S} (S_{b(w)})^{\tau_{a\rightarrow b}P} \right) \cap S[b; P],
\]

where, recall from above, \( S[b; P] \) is the product order on \( S[b; P] \), and

\[
S_b[b; P] := S_b \cap S[b; P].
\]

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 13. 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
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
Chapter 2. Trust, Distrust, Semi-Trust

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. 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-

---

ally aim at persuading hearers w.r.t. certain propositions, getting them to adopt the same attitude towards the proposition she already has herself.\footnote{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, Sietsema, and Wang 2012). Persuasiveness is also studied in the multi-agent systems literature (Dunin-Keplicz and Verbrugge 2010) and in argumentation theory (Walton and Krabbe 1995).}

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 \(\mathcal{C} = (\{S_a\}_{a \in A}, T)\), an introspective propositional attitude \(A\), a proposition \(P\), an agent \(a \in A\) and a world \(w \in S\), let us write

\[
\mathcal{C}, w \models A_a P \iff S_{a(w)} \models AP.
\]

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 \(\mathcal{C}, w \models A_a P\).

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

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

\[
\mathcal{C}, w \models B_a P.
\]

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 \(S_{a(w)}\) 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

\[
\text{if } T_w(b, a) = \tau, \text{ then } \mathcal{C}, w \models \neg \tau_a P.
\]

We say that \([a : P]\) is honest in \(\mathcal{C}\) at \(w\) iff \([a : P]\) is honest towards every \(b \neq a \in A\) in \(\mathcal{C}\) at \(w\).
Chapter 2. Trust, Distrust, Semi-Trust

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 \( C, w \models B_aP \).

Since dynamic attitudes create their fixed point, we have, assuming that \( w \in S[a: P] \), that \( C[a: P], w \models BP \). 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 \([a: P]\text{ is honest towards } b \text{ at } w \text{ in } C \iff C, w \models K_aP\).
2. If \( T_w(b, a) = \uparrow^+, \text{then } [a: P]\text{ is honest towards } b \text{ at } w \text{ in } C \iff C, w \models Sb_aP\).
3. If \( T_w(b, a) = \uparrow^+, \text{then } [a: P]\text{ is honest towards } b \text{ at } w \text{ in } C \iff C, w \models B_aP\).

**Proof.** We show the first item. Suppose that \( T_w(b, a) = ! \). Then \([a: P]\text{ is honest iff } C, w \models \overline{1}_aP\). But since \( \overline{1} = K \), the latter is the case iff \( C, w \models K_a\varphi \).

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 $\tau$, the fixed point of $\tau$ entails the opposite of belief $B\neg$, hence for $[a: P]$ to be honest, it is required that $C, w \vDash B\neg P$. But since $C, w \vDash KP$ by our assumption, it is not the case that $C, w \vDash B\neg P$, so $[a: P]$ is not honest.
On our formalization, then, \([a: p]\) constitutes a “sincere lie” in \(C\) 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 \([a: \neg P]\), on the other hand, is honest, but not sincere in \(C\) 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 \(\S 2.8.8\) 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 \(\hat{\tau}^+\), then \([a: P]\) is honest towards \(b\) at \(w\) iff \([a: P]\) 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 \([a: P]\) that is honest at \(w\) in \(C\) is also sincere at \(w\) in \(C\).

To see this, suppose that \(T_w(b,a) = \tau\) is strictly positive, and assume that \([a: \phi]\) is honest towards \(b\) at \(w\) in \(C\). By definition of honesty, \(C, w \models \tau_a P\). But our earlier results imply that \(\tau \leq B\) for any strictly positive \(\tau\) (Proof: Suppose \(\tau\) is strictly positive. By Proposition 32, \(\tau \leq \tau^+\). By Proposition 11, \(\tau \leq \tau^{\hat{\tau}^+}\), and since \(\hat{\tau}^+ = B\), it follows that \(\tau \leq B\).) Thus, since \(C, w \models \tau_a P\), it follows that \(C, w \models B_a P\), so \([a: P]\) is sincere at \(w\) in \(C\).

It is not, however, in general the case, that a sincere communication act is honest towards a hearer who has a strictly positive attitude 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 \(\hat{\uparrow}^+\) towards the speaker \(a\), then the speaker needs to strongly believe that \(P\) for the communication act \([a: P]\) 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 $C$ at $w$ iff

$$\text{if } C, w \models B_a Q, \text{ then } C[a: P], w \models B_b Q.$$ 

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 $S_b$ (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 $C$ and $w$ are such that

1. $T_w(b, a) = \uparrow^+$,
2. $C, w \models B_a Q$,
3. $C, w \models \neg S_b a Q$.
4. $C, w \models \neg B_b Q$.

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.
Our question is: is there a communication act \([a : P]\) which is honest, sincere and persuasive w.r.t. \(Q\) towards \(b\) in \(C\) at \(w\)? Spelled out in more detail, \([a : P]\) meets our requirements iff

1. \(C, w \models B_a P\) (required by sincerity),
2. \(w \models C Sb_a Q\) (required by honesty), and
3. \(w \models C[a : P] B_b Q\) (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, \(C, w \not\models Sb_a Q\).
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

\[
P := \{ v \in S \mid C, v \models K_a Q \}.
\]

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

\[
P := \{ v \in S \mid C, v \models B_a Q \}.
\]

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 \(C\) yields just \(C\): 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 \textit{defeasibly knows that} \( Q \), that is, choose \( P \) by means of

\[
P := \{ v \in S \mid C, v \models \Box_a Q \}.
\]

Let us verify that, under this choice of \( P \), \([a : P]\) is sincere, honest and persuasive towards \( b \) in \( C \) at \( w \) w.r.t. \( Q \).

— \textit{Sincerity:} Observe that for any trust-plausibility order \( C' \) and world \( v \in S' \): \( C', v \models B_a Q \) iff \( C', v \models B_a \Box_a Q \). 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 \( C, w \models B_a Q \), \([a : Q]\) is in fact sincere, so the same goes for \([a : P]\).

— \textit{Honesty:} for any trust-plausibility order \( C' \) and world \( v \in S' \): \( C', v \models B_a Q \) iff \( C', w \models S_b \Box_a Q \). Since, by our assumption, \( C, w \models B_a Q \), it follows that \( C, w \models S_b \Box_a Q \). For \([a : Q]\) to be honest, on the other hand, it is required that \( C, w \models S_b P \) (since \( T_w(b, a) = \Uparrow^+ \)), but this is just what we have verified, so \( C, w \models S_b P \). So \([a : P]\) is honest.

— \textit{Persuasiveness:} To see that \([a : P]\) is persuasive, notice that in \( C[a : P] \), agent \( b \) will consider all worlds in \( P \cap S[a : P] \) strictly more plausible than all other worlds. This implies that \( C[a : P], w \models B_b \Box_a Q \). But notice that for any trust-plausibility order \( C' \) and world \( v \in S' \): if \( C', v \models \Box_a Q \), then \( C', v \models Q \). Hence \( C[a : Q], w \models B_b Q \). 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 \textit{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.