Doxastic Attitudes as Belief-Revision Policies

Alexandru Baltag Ben Rodenhäuser Sonja Smets
ILLC, University of Amsterdam

Abstract. While propositional doxastic attitudes, like knowledge and belief, capture an agent's opinion about certain propositions, her attitudes towards *sources* of information express her opinion about the reliability (or trustworthiness) of those sources. If an agent trusts a witness, then she will, within certain limits, tend to accept his testimony as veridical. But if she considers the witness to be a notorious liar, she may come to believe the opposite of what he tells her. In this paper, we put such attitudes towards sources (or *dynamic (doxastic) attitudes*) center stage, and formalize them as *belief-revision strategies*: policies governing how an agent changess her beliefs whenever new information from a certain (type of) source is received. We present a semantic, qualitative modelling of this notion and investigate its properties.

Introduction

This paper explores the idea that an agent's "information uptake" (i.e. what she does with some new informational input) depends substantially on her *attitude* towards the source of information: her assessment of the reliability of the source. Evidence obtained by direct observation, e.g., is normally considered to be more reliable than testimonial evidence, and testimonies from different witnesses may be differently assessed, depending on the reliability of each witness. Formally, we encode attitudes towards sources as *strategies for belief change*, applicable to any information received from a particular (type of) source. Such a strategy *pre-encodes* what the recipient will do if an input from that source is received.

Traces of this idea are found scattered across the literature, sometimes formulated in terms of notions like "evidential reliability" or "epistemic trust".[1] In Bayesian Epistemology, the reliability of a source is captured by weights or probabilities attached to the new information, which determine different ways of processing it (e.g., Bayesian conditioning versus Jeffrey conditioning).[2] In Belief Revision theory, various methods for (iterated) belief revision have been proposed that can be understood as corresponding to different attitudes to the incoming information.[3]

While most previous authors have focused on quantitative approaches formalizing *degrees of acceptance* or *degrees of trust*, we propose a *qualitative-relational setting* that allows us to model a *much more general class of doxastic attitudes*, including various forms of trust, distrust and "semi-trust". We give a semantic formalization of these concepts, study the *strength order* between different dynamic attitudes and the *natural operations* with them, and show how the standard propositional attitudes can be recovered as *fixed points* of dynamic attitudes. In the

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

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

[3] Boutilier (1996), Spohn (1985, 2009), Nayak (1994), Rott (2004, 2006), among others.

conclusion, we discuss a multi-agent extension of the setting, in which various properties of communication acts, like honesty and sincerity, can be studied.

1 Background

We briefly review a fairly standard setting building on prior work in (semantically oriented) Belief Revision theory and Dynamic Epistemic Logic.[4] The main notions are *plausibility orders*, *upgrades* (i.e., order transformers) and *propositional (doxastic) attitudes.*

Plausibility Orders. Fix a countable set Σ, called *the set of all possible worlds* (or *possible states* of the world). A *proposition* is a set of possible worlds. A *plausibility order* is a pair $\mathcal{S} = (\mathrm{S}, \leq)$, where $\mathrm{S} \subseteq \Sigma$ is finite, and $\leq \subseteq \mathrm{S} \times \mathrm{S}$ is a total preorder, i.e., a transitive, connected (and thus reflexive) relation.

The fact that $s \leq t$ indicates that state s is *at least as plausible as* state t, from the perspective of our agent. We write $\mathrm{best}_{\mathcal{S}}\, P$ for the *most plausible P-states* in the order $\mathcal{S}$, given by $\mathrm{best}_{\mathcal{S}}\, P := \{s \in P \mid \forall t \in P : s \leq t\}$. We write $\mathrm{best}\, \mathcal{S}$ for $\mathrm{best}_{\mathcal{S}}\, \mathrm{S}$.

The set of all possible worlds Σ comprises the totality of all possibilities that are consistent with some unchangeable, context-independent and time-independent information about the world. In addition, an agent may gain, over time, more information about the world, thereby reducing the initial set of possibilities to a smaller set, embodying her *hard information*, assumed to be absolutely certain, i.e., irrevocably known by the agent. This hard information is what the domain S of a plausibility order $\mathcal{S}$ captures.

Going further, the agent may also possess *soft information*, that is *not* absolutely certain but *subject to revision*, and that merely allows her to hierarchize the possibilities consistent with her hard information according to their subjective "plausibility", but not to exclude any of them. This relative hierarchy is represented by the relation $\leq$.

Propositional Attitudes. Plausibility orders allow us to capture a variety of "opinions" an agent might have about a given proposition. A *(doxastic) propositional attitude* is a function

$$A : \mathcal{S}, P \longmapsto A_{\mathcal{S}}P$$

that assigns a proposition $A_{\mathcal{S}}P \subseteq \mathrm{S}$ to each given plausibility order $\mathcal{S}$ and proposition P. Important examples of such attitudes are:

- *(Irrevocable) knowledge K*, defined by $K_{\mathcal{S}}P := \{s \in \mathrm{S} \mid \mathrm{S} \subseteq P\}$.
- *(Simple) belief B*, defined by $B_{\mathcal{S}}P := \{s \in \mathrm{S} \mid \mathrm{best}\, \mathcal{S} \subseteq P\}$.
- *Strong belief Sb*, defined by $Sb_{\mathcal{S}}P := \{s \in \mathrm{S} \mid P \cap \mathrm{S} \neq \varnothing \wedge \forall t \in P \forall r \notin P : t < r\}$.
- *Triviality* $\top$, defined by $\top_{\mathcal{S}}P := \mathrm{S}$.
- *Inconsistency* $\bot$, defined by $\bot_{\mathcal{S}}P := \varnothing$.

Upgrades. We may now wonder how to represent belief changes in the setting given by plausibility orders. A *(doxastic) upgrade* u is a function

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

that takes a given plausibility order $\mathcal{S} = (\mathrm{S}, \leq)$ to a plausibility order $\mathcal{S}^u := (\mathrm{S}^u, \leq^u)$, satisfying that $\mathrm{S}^u \subseteq \mathrm{S}$.

The *composition* $u \cdot u'$ of two upgrades u and u' is given as usual, by $\mathcal{S}^{u \cdot u'} = (\mathcal{S}^u)^{u'}$, i.e., "first apply u, then apply u'."

[4] Settings of this kind, or close relatives, are discussed, in various degrees of generality, by many authors, cf., e.g., Spohn (1985), Grove (1988), Boutilier (1996), van Benthem (2007), Baltag and Smets (2008).

Important examples of upgrades include the following.[5] For each proposition P,

- the *update* $!P$ maps each plausibility order $\mathcal{S}$ to the relativization of the order with P, i.e., all non-P-states are deleted, everything else is kept the same.

- the *radical upgrade* $\Uparrow P$ makes all P-states more plausible than all non-P-states, leaving everything else unchanged.

- the *positive radical upgrade* $\Uparrow^+ P$ is defined exactly as radical upgrade $\Uparrow P$, except that, for any proposition P and order $\mathcal{S}$ such that $P \cap \mathrm{S} = \varnothing$, the result of applying $\Uparrow^+ P$ to $\mathcal{S}$ is the empty plausibility order.

- the *conservative upgrade* $\uparrow P$ promotes the best P-states (i.e., the states in $\text{best}_{\mathcal{S}} P$) to become the best states overall, leaving everything else unchanged.

- the *positive conservative upgrade* $\uparrow^+ P$ is defined exactly as conservative upgrade $\uparrow P$, except that, for any proposition P and order $\mathcal{S}$ such that $P \cap \mathrm{S} = \varnothing$, the result of applying $\uparrow^+ P$ to $\mathcal{S}$ is the empty plausibility order.

- the *semi-positive conservative upgrade* $\uparrow^{\sim +} P$ adds the best P-states to the best states overall, leaving everything else unchanged.

- the *null upgrade* $\varnothing$ maps every plausibility order to the empty plausibility order.

- the *trivial upgrade* id maps every plausibility order to itself.

2 Dynamic Doxastic Attitudes

A *dynamic (doxastic) attitude* τ is a function

$$\tau : P \longmapsto \tau P$$

that maps each proposition P to an upgrade τP, satisfying

1. $\mathcal{S}^{\tau P} = \mathcal{S}^{\tau(P \cap \mathrm{S})}$,
2. $\tau P \cdot \tau P = \tau P$,
3. if $P \in \{\varnothing, \Sigma\}$, then $\tau P \in \{\varnothing, id\}$.

The *first condition* says that the result of applying τP does not depend on the worlds satisfying P that are *outside* of S. The *second condition* says that dynamic attitudes are *idempotent* if their propositional argument is kept fixed: receiving the very same semantic information one has just received is redundant. The *third condition* says that dynamic attitudes deal in a uniform manner with information that is trivial or inconsistent: such information uniformly leaves the order unaffected ($\tau P = id$) or deletes the whole order ($\tau P = \varnothing$).

We have said that we understand dynamic attitudes as strategies for belief change. We can now spell out such a strategy from the perspective of an agent as follows:

> "Whenever I receive the information that P from a source towards which I have the attitude τ, I will apply the upgrade τP to my current plausibility order."

Our above examples of upgrades readily translate to examples of dynamic attitudes:

- *infallible trust* ! maps each proposition P to the update $!P$;

- *strong trust* $\Uparrow$ (resp. *strong positive trust* $\Uparrow^+$) maps each P to the radical upgrade $\Uparrow P$ (resp. positive radical upgrade $\Uparrow^+ P$);

[5] For a more thorough discussion of upgrades, see, e.g., Baltag and Smets (2008).

- *minimal trust* $\uparrow$ (resp. *positive minimal trust* $\uparrow^{+}$) maps each P to the conservative upgrade $\uparrow P$ (resp. positive conservative upgrade $\uparrow^{+} P$);
- *semi-positive minimal trust* $\uparrow^{\sim +}$ maps each P to the semi-positive conservative upgrade $\uparrow^{\sim +} P$;
- *neutrality id* maps each P to the identity upgrade *id*;
- *isolation* $\varnothing$ maps each P to the null upgrade $\varnothing$.

Positive Attitudes. Within our framework, we can capture a notion of "acceptance" or "trust" (as discussed in the introduction) in the following way. An attitude τ is *positive* if it satisfies:

1. $P \cap \mathrm{S} \neq \varnothing \implies \mathrm{S}^{\tau P} \neq \varnothing$;
2. $\mathrm{best}\, \mathcal{S}^{\tau P} \subseteq P$.

Essentially, having a positive attitude towards a source means that the agent *will always come to believe any information received from that source*, and moreover *her new beliefs will be consistent whenever this is possible* (i.e. whenever the new information is consistent with her prior knowledge).

Among the attitudes in our list of examples above, the only positive ones are infallible trust !, strong positive trust $\Uparrow^{+}$ and minimal positive trust $\uparrow^{+}$.

Weakly Positive Attitudes. A dynamic attitude τ is *weakly positive* if

$$P \cap \mathrm{S} \neq \varnothing \quad \implies \quad \mathrm{S}^{\tau P} \neq \varnothing \wedge \mathrm{best}\, \mathcal{S}^{\tau P} \subseteq P.$$

Essentially, having a weakly positive attitude towards a source means that, *as long as the new information received from that source is consistent with the agent's prior knowledge, her new set of beliefs will be consistent and include the new information.*

As the terminology suggests, all positive attitudes are weakly positive. In fact, all our examples of dynamic attitudes so far, except for neutrality, isolation and semi-positive minimal trust, are weakly positive.

Observe that (weakly) positive attitudes capture a type of *uniform trust*: the agent processes in the same (positive) way every information coming from this source. In Section 5 we'll model more realistic, contextually-dependent forms of trust.

Semi-Positive Attitudes. Many people regard the weather forecast as a source of positive, but inconclusive evidence. If the forecast predicts sun, they are still inclined to take their umbrella; on the other hand, they also take a light shirt. The weather forecast leads them to drop their belief that it will rain, without their coming to believe the opposite. This is an example of a "semi-trusting" attitude, in-between trust and distrust. Formally, an attitude τ is *semi-positive* if

$$P \cap \mathrm{S} \neq \varnothing \quad \implies \quad P \cap \mathrm{best}\, \mathcal{S}^{\tau P} \neq \varnothing.$$

In this way, we arrive at a class containing all the weakly positive attitudes, but also further interesting ones, such as *semi-positive minimal trust* $\uparrow^{\sim +}$.

(Weakly, or semi-) negative attitudes. These can be defined by dualizing the above clauses (though a shorter definition will be given in Section 5). Weakly negative attitudes formalize "uniform distrust", i.e., the source is generally mis-trusted, independent on the content of the new information, and hence our agent essentially upgrades "to the contrary".

3 Fixed Points

Intuitively, an upgrade τP is *redundant*, or *uninformative*, in an order $\mathcal{S}$ if the order remains unchanged when applying τP, i.e., if $\mathcal{S}^{\tau P} = \mathcal{S}$. Such "uninformativity" of a given (dynamic-doxastic) attitude τ is itself a *propositional* (doxastic) attitude $\overline{\tau}$.

Formally, given a dynamic attitude τ, the *fixed point* $\overline{\tau}$ of τ is the propositional attitude $\overline{\tau} : \mathcal{S}, P \mapsto \overline{\tau}_{\mathcal{S}} P$, defined by

$$\overline{\tau}_{\mathcal{S}} P \quad := \quad \{s \in \mathrm{S} \mid \mathcal{S}^{\tau P} = \mathcal{S}\}.$$

Intuitively, $\overline{\tau} P$ holds whenever the agent *has "already learned" everything she could learn from a P-asserting, τ-trusted source.* This notion allows us to link dynamic and propositional attitudes in a perspicuous way:

Proposition 1.

- *The fixed point of infallible trust is knowledge:* $\overline{!} = K$.
- *The fixed point of strong positive trust is strong belief:* $\overline{\Uparrow^+} = Sb$.
- *The fixed point of positive minimal trust is belief:* $\overline{\uparrow^+} = B$.
- *The fixed point of neutrality is triviality:* $\overline{id} = \top$.
- *The fixed point of isolation is inconsistency:* $\overline{\varnothing} = \bot$.

4 The Strength Order

It is natural to compare propositional attitudes to each other, e.g., knowledge K is *stronger than* belief B, since knowledge *entails* belief.

We introduce an analogue notion of strength for dynamic attitudes. Given dynamic attitudes σ and τ, the attitude σ is *at least as strong as* τ (notation: $\sigma \leq \tau$) iff the following holds:

$$\sigma \leq \tau \quad \Longleftrightarrow \quad \forall P \; (\sigma P \cdot \tau P = \sigma P)\,.$$

Intuitively, this relation of strength captures that a source towards which the agent has the attitude σ "subsumes" a source towards which the agent has the attitude τ: after the first source tells you that P, you don't need to "hear it again" from the second source. The strength relation on dynamic attitudes matches the entailment relation on their fixed points familiar from epistemic logic:

Proposition 2. *For all dynamic attitudes σ and τ:*

$$\sigma \leq \tau \quad \Longleftrightarrow \quad \forall \mathcal{S} \forall P \; (\overline{\sigma}_{\mathcal{S}} P \Longrightarrow \overline{\tau}_{\mathcal{S}} P)\,.$$

The strength order is bounded by isolation and neutrality, while infallible trust and minimal trust emerge as strongest, resp. weakest weakly positive attitudes:

Proposition 3.

- *For all dynamic attitudes τ:* $\varnothing < \tau < id$.
- *For all weakly positive attitudes τ:* $! \leq \tau \leq \uparrow$.
- *For all positive attitudes τ:* $! \leq \tau \leq \uparrow^+$.
- *For all semi-positive attitudes τ:* $! \leq \tau \leq \uparrow^{\sim +}$.

5 Operations with Attitudes

We present here four natural operations with dynamic doxastic attitudes.

1. Opposite. The *opposite* $\tau^\neg$ of an attitude τ is given by $\tau^\neg P := \tau(\neg P)$. Using this notion, we obtain attitudes like *minimal negative distrust* $\uparrow^- := (\uparrow^+)^\neg$ (the opposite of positive minimal trust), *strong distrust* $\Uparrow^\neg$ (the opposite of strong trust) etc. More generally, *(weakly, or semi-) negative attitudes* can now be defined as the attitudes τ of the form $\tau = \sigma^\neg$, for some (weakly, or semi-) positive attitude σ.

2. Contextual Mixtures. Whether a source is trusted (in a certain way, as given by a particular weakly positive attitude) may depend on the topic of conversation.

E.g., a certain source may be trusted if she is making mathematical statements, but not on other topics. We capture such finer distinctions as follows. A *context* is a set of propositions. Given two dynamic attitudes σ and τ and a context Γ, the *mixture* $\sigma_\Gamma\tau$ of σ and τ w.r.t. Γ is given by

$$\sigma_\Gamma\tau P := \begin{cases} \sigma P & P \in \Gamma \\ \tau P & P \notin \Gamma \end{cases}$$

In this way, we can "mix" positive and negative attitudes, or weakly positive and semi-positive ones etc.

3. Restrictions. A special case of a mixture is the *restriction of an attitude τ to a context* Γ, given by

$$\tau|_\Gamma := \tau_\Gamma\varnothing$$

4. Implicative Closure. If a speaker, when asked whether she will come to a party, says *I have to work*, we typically understand that she will *not* come: for if both were true—that she has to work, and that she will come—, it would have been more common to simply say *I will come to the party.*

We capture such "Gricean" phenomena as follows. An *implicative context* is a strict partial order $<$ on $\wp(\Sigma)$, the set of propositions. If $Q < P$, then we interpret this as: "if P and Q are the case, then it is more common to say Q than P". Using the notion of *implicative closure of a proposition* P (w.r.t. an implicative context $<$), given by

$$P^< := P \setminus \bigcap_{Q<P} Q\,,$$

we define the *implicative closure $\tau^<$ of an attitude* τ (w.r.t. implicative context $<$) by

$$\tau^< P \;:=\; \tau(P^<)\,.$$

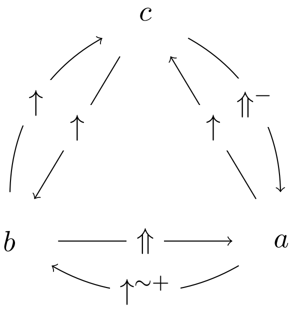
In this way, we can capture *default rules of communication*, i.e., upgrades based on certain assumptions about "when people normally say certain things".

Conclusions and Future Work

Our setting can be extended to an analysis of information exchange among communicating agents. In a communication setting, sources of information are *speakers* making assertions, and attitudes towards sources become attitudes of *hearers towards speakers.*

To capture such scenarios, we introduce plausibility frames, which are just the the natural multi-agent versions of our notion of plausibility order.[6] For a set of agents $\mathcal{A}$, a *plausibility frame* $(\mathrm{S}, \sim_a, \leq_a)_{a\in\mathcal{A}}$, is a set of states $\mathrm{S} \subseteq \Sigma$, together with an agent-indexed family of equivalence relations $\sim_a$ and relations $\leq_a$ such that $\leq_a$ respects $\sim_a$ (i.e. $s \leq_a t$ implies $s \sim_a t$), and the restriction of $\leq_a$ to each $\sim_a$-equivalence class is a plausibility order (in the sense of Section 1).

We represent the agents' mutual attitudes towards each other at a given state by means of a *trust graph*, whose nodes are agents, and whose edges from any node a to any other node b are labeled with the attitude of agent a towards agent b. E.g. in the trust graph depicted below, agent a has an attitude of minimal trust $\uparrow$ towards agent c:



Given a plausibility model together with a trust graph at each state, the effect of an assertion of P made by speaker a on the belief state of hearer b is given by upgrading the relation $\leq_b$ within each $\sim_b$-equivalence class according to b's attitude towards a.[7]

[6] cf. Baltag and Smets (2008), van Benthem (2007).

[7] We assume that agents are *introspective w.r.t.*

The interest of this multi-agent version derives from the fact that it now becomes possible to investigate *properties of assertions* made by a speaker, building on the work of the previous sections. For instance, an assertion of P is *sincere* if the speaker *believes that* P. The assertion is called *honest* if the *fixed point* of the hearer's dynamic attitude towards the speaker is the same as the speaker's propositional attitude towards P: this means that the hearer's attitude towards the speaker is *matched* by the speaker's attitude to her assertion. If, e.g., the hearer's attitude towards the speaker is *positive minimal trust* $\uparrow^+$, honesty requires that the speaker actually *believes that* P (since the fixed point of $\uparrow^+$ is belief).

One interesting observation is that *distrust induces a tension between sincerity and honesty.* For example, consider the reply given by Brad Pitt to *Wired* magazine about people lying (in their online dating profile) on how much money they earn:

> *Everyone lies online. In fact, readers expect you to lie. If you don't (exaggerate your income), they'll think you make less than you actually do. So the only way to tell the truth is to lie.*

We can capture an analogue of this phenomenon: suppose it is common knowledge that the hearer *distrusts* (has attitude $\uparrow^-$ towards) the speaker. If the speaker believes P, then asserting P would be *sincere* but *dishonest*: indeed, the speaker well knows that this would only lead the hearer to believe $\neg P$. So, if the speaker really wants to "convert" the hearer to her own propositional attitude (of belief) towards P, she should rather assert the *opposite* of what she believes, i.e. $\neg P$. Asserting $\neg P$ in this scenario is an example of an "honest lie", aimed at getting the hearer to adopt what the speaker believes to be the "correct" attitude towards the issue in question.

In on-going work, we develop this setting into a general study of the *formal epistemology of testimony, persuasion and trust.*

their own attitudes, so that b's attitude towards a is the same in all the states that belong to the same $\sim_b$-equivalence class.

References

Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. *Texts in Logic and Games*, 3, 2008.

Craig Boutilier. Iterated revision and minimal change of conditional beliefs. *Journal of Philosophical Logic*, 25(3):263–305, 1996.

Adam Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988.

Joseph Halpern. *Reasoning about Uncertainty.* MIT Press, 2003.

Richard Jeffrey. *Subjective Probability, The Real Thing.* Cambridge University Press, 2004.

Keith Lehrer and Carl Wagner. *Rational Consensus in Science and Society.* Reidel, 1981.

Abhaya C Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41: 353–390, 1994.

Hans Rott. Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, 61(2-3):469–493, 2004.

Hans Rott. Shifting priorities: Simple representations for twenty-seven iterated theory change operators. 2006.

Wolfgang Spohn. Ordinal conditional functions: A dynamic theory of epistemic states. In William L. Harper and Bryan Skyrms, editors, *Proceedings of the Irvine Conference onf Probability and Causation*, volume II. 1985.

Wolfgang Spohn. A survey of ranking theory. In Franz Huber and Christoph Schmidt-Petri, editors, *Degrees of Belief.* Springer, 2009.

Johan van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 14:129–155, 2007.