Joint Trust for Belief Revision

Aaron Hunter [†,*], Richard Booth[‡]
[†] BC Institute of Technology, Burnaby, Canada
[‡] Cardiff University, Cardiff, Wales

Abstract

The process of belief revision is impacted by trust relationships between agents. In the simplest case, information is reported from a single source and the belief change that occurs is dependent on the extent to which that source is trusted over a particular domain. In this paper, we are concerned with the more complicated case where the new information is reported by a set of agents. We first introduce a simple model of trust in an agent, and show how it influences the process of belief revision. We then define an joint notion of trust that is built by combining the trust held in each individual agent in the reporting set. We use this formal framework to define precisely when a collection of agents can be seen as a trusted authority over a particular formula for revision. While our framework is based on a particular model of trust, we argue that this approach can be used to define a suitable notion of joint trust in a wide range of settings.

Keywords: belief revision, trust, multi-agent systems

1. Introduction

We are concerned with the manner in which trust impacts belief revision. In traditional approaches to belief revision, the focus is on determining which of the initial beliefs are easiest to abandon when there is a conflict with the new information. The underlying assumption is that the new information is, in some sense, more reliable than the initial beliefs. In practice, this is not always a reasonable assumption. Information is often obtained from another agent or set of agents. The extent to which new information is incorporated depends on the trust held in the reporting agents, with respect to a particular domain. In this paper, we set out to formally define a notion of joint trust and the impact of this form of trust on belief revision.

This paper makes several contributions to the existing literature on trust and belief revision. First, we provide a precise definition of what it means to say that an agent A trusts another agent B on a formula ϕ. Second, we define a formal notion of *joint trust* with respect to a finite set of agents. This notion is based on a simple algebra that allows any set of agents to be combined to a single joint agent that is equivalent with respect to our notion of trust.

2. Preliminaries

2.1. Motivation

We focus on belief change and knowledge-based trust. That is, we are interested in the manner in which an agent is trusted based on what they are *perceived to know*. This is distinct from trust related to honesty, where we may discount reports from an agent that is lying [1]. Knowledge-based trust has the property that an agent might be trusted when they assert a formula ϕ, while they are not trusted if they assert $\neg\phi$. Similarly, there are cases where each agent in a group is not trusted on a particular statement, but the group is jointly trusted over the same statement. We illustrate these two points with the following motivating example.

[*]aaron_hunter@bcit.ca

Example 1. *Suppose that there are agents A, B and C responsible for observing a sick animal at a zoo over a period of two days. Informally, A is a manager that does not interact with the animal directly, whereas B and C are tasked with observing its behaviour. On day 1, B watches the animal all day. On day 2, C watches the animal all day. Consider three possible things that A might be told after two days of observation.*

(1) C says "The animal has eaten." (E)
(2) C says "The animal has not eaten." ($\neg E$)
(3) B and C jointly say "The animal has not eaten." ($\neg E$)

It seems that A should believe the first statement from C, because C was in a position to observe the animal eating. However, this is not the case for the second statement. Since C did not observe the animal for the entire duration, C can not be certain that the animal has not eaten. Hence, C is trusted over the statement E but not over the statement $\neg E$.

The third statement should again be believed, as B and C together have observed the animal for the entire duration. Hence, although $\neg E$ would not be believed if reported by either B or C individually, it should be believed if they report it jointly.

We are interested in defining a framework for reasoning about trust in this kind of example, where reports can come either from an individual or from a group. For instance, we can see that C should be trusted over E but not over $\neg E$. However, when we consider joint announcements, we can see that the set of agents $\{B, C\}$ should be trusted over $\neg E$. Hence, there is a natural sense in which trust in individuals can be combined to determine the kind of trust held in a group.

In this paper, we formalize these notions of trust in terms of how they impact the process of *belief revision*. For example, when C reports the formula E, how should A revise their beliefs? Certainly not in the same manner that they would revise their beliefs if C were to report $\neg E$, based on our informal discussion of trust.

2.2. **Belief Revision**

Let $\mathcal{L}$ be a finite propositional vocabulary. A *sentence* (or *formula*) over $\mathcal{L}$ is constructed using the propositional connectives $\{\wedge, \vee, \neg\}$. A *state* is a subset of $\mathcal{L}$ representing the set of atoms that are true. The notion of truth in a state (written $s \models \phi$) is defined in the usual manner of propositional logic, as is the notion of propositional entailment ($\psi \models \phi$).

The dominant approach to logic-based belief revision is the AGM approach [2]. We briefly introduce the framework here. A *belief set* K is a logically closed set of formulas. An AGM belief revision operator $*$ is a function that maps a belief set K and a formula ϕ to a new belief set $K * \phi$. An AGM belief revision operator is further constrained in that it must satisfy the AGM postulates for revision. In the interest of space, we do not write the postulates here. However, it is well-known that every AGM belief revision operator can be defined in terms of a *faithful ordering* $\prec$ over states [3]. The essential feature of AGM revision operators is that $K * \phi$ is the set of formulas that are true in the $\prec$-minimal states consistent with ϕ.

2.3. **Trust-Sensitive Belief Revision**

One of the AGM postulates is the *success postulate*, which asserts that $K * \phi \models \phi$. This postulate constrains belief revision so that new information is always believed. While this is sensible when the new information is seen as infallible, it is not sensible when the new information is reported by an agent that might be mistaken. In order to address this problem in practice, we need to add a model of trust on top of our belief revision operator.

One approach to adding trust is through so-called *trust-sensitive* belief revision operators. In this approach, a partition over states is associated with each reporting agent. Informally,

an agent is not trusted to be able to distinguish between states in the same cell of the partition. Using this partition, we can define a trust-sensitive revision operator $*$ that only incorporates the part of a formula ϕ over which the reporting agent can be trusted. This operator can be characterized by a set of postulates, provided in [4].

There have been other models of trust in related settings. For example, the interaction between trust and belief change has been modelled in Dynamic Epistemic Logic [5, 6]. The notion of expertise has also been modelled in an extension of propositional logic [7]. However, the work in this paper is most closely related to trust-sensitive revision, as we are interested in modeling the notion of trust as it relates to an existing belief revision operator.

3. Defining Trust

3.1. Trust Scenarios

Assume a fixed propositional vocabularly $\mathcal{L}$. We are interested in scenarios involving a set of agents that trust each other over different topics. In order to define the scenario of interest more precisely, we need to define the notion of a *trust operator*. In the following definition, let $sent(\mathcal{L})$ denote the set of propositional sentences over $\mathcal{L}$.

Definition 1. *Given a set* $\mathbf{A}$ *of agents, a trust operator is a function* $S : \mathbf{A} \to 2^{sent(\mathcal{L})}$.

A trust operator S maps each agent to a set of propositional sentences. Informally, an agent is only trusted when they report a sentence $\phi \in S(A)$. We are intentionally defining trust operators in a very general setting at this point, so the trust set need not even be closed under logical consequence.

We can now formally define a trust scenario.

Definition 2. *A* trust scenario *over* $\mathcal{L}$ *is a triple* $\langle \mathbf{A}, R, S \rangle$ *where*

- $\mathbf{A}$ *is a set of* agents.
- R *is a function that maps each* $A \in \mathbf{A}$ *to an AGM revision operator* $*_A$.
- S *is a trust operator that maps each agent* B *to the set of sentences* $S(B)$.

The revision function $*_A$ is the idealized revision operator that the agent A would use to incorporate new information from a perfectly reliable source. The function S encodes the areas of expertise of the other agents; so each agent should only be believed when they assert a sentence in $S(B)$. When A receives new information, then $*_A$ and $S(B)$ must be used together to determine what A should believe.

3.2. Revision by Reports

In a trust scenario, all new information must have an associated *source*. Hence, we will talk about an agent A receiving a *report* of ϕ from an agent B; we represent reports as ordered pairs of the form (B, ϕ). This simply means that the formula ϕ has been provided to A by the agent B. The way the information is incorporated by A will be influenced by the trust held in that set.

We would like to define a suitable approach to revision by a report, but we first motivate the solution with a simple example. Suppose that we have a trust scenario where $p \in S(B)$ but $p \wedge q \notin S(B)$. In this case, what should A believe after the report $(B, p \wedge q)$? Since the source B is not trusted over $p \wedge q$, A should not just revise by this formula. However, since $p \wedge q$ implies p, we can see that B is actually asserting p is true. Since A trusts B over p, it follows that p should be believed after revision. We therefore would like to revise by the 'strongest' formula in $S(B)$ that is logically entailed by the report.

We formalize the reasoning in the preceding example in the following definitions. In these definitions, we call ψ a *weakening* of ϕ just in case $\phi \models \psi$. Similarly, we call ψ a *strengthening* of ϕ just in case $\psi \models \phi$.

Definition 3. *Let ϕ be a formula, and let Ψ be a set of formulas. We call ψ a* strongest weakening *of ϕ with respect to Ψ just in case*

(1) $\psi \in \Psi$.
(2) ψ is a weakening of ϕ.
(3) For all weakenings $\psi' \in \Psi$ of ϕ, $\psi \models \psi'$.

We write $\phi[\Psi]$ as a shorthand for the strongest weakening of ϕ with respect to Ψ.

Using this definition, we can define a form of revision for reports.

Definition 4. *Let $\langle \mathbf{A}, R, S\rangle$ be a trust scenario. For any belief set K, agents $A, B \in \mathbf{A}$ and ϕ, define:*

$$K *_A (B, \phi) = \begin{cases} K *_A \phi[S(B)], & \text{if } S(B) \neq \emptyset \\ K, & \text{otherwise} \end{cases}$$

Note that we are overloading the $*_A$ operator in this definition. We are actually defining a new operator that takes reports as inputs on the left hand side, but we are giving the result in terms of the original $*_A$ on the right hand side. It will always be clear from the argument what $*_A$ stands for. The important point of the definition is that, when B reports the formula ϕ to the agent A, then A revises their beliefs by the strongest weakening of ϕ over which B is considered an authority.

Example 2. *We can define a trust scenario over the language $\{E\}$ for our motivating example. For the moment, we only specify the revision operator and trust function for A:*

- *$*_A$ is the Dalal revision operator[8].*
- *$S(B) = S(C) = \{E\}$.*

Note that we would use any AGM operator, we only specify the Dalal operator to have something concrete. Now suppose that the initial belief set of A is $K = \emptyset$. This means that A does not believe either E or $\neg E$ at the outset. It is easy to verify the following:

- *$K *_A (C, E) \models E$.*
- *$K *_A (C, \neg E) \not\models \neg E$.*

This is the desired result for the case of single-agent reports.

The model of trust being used here is actually a generalization of the partition-based model of [4].

Proposition 1. *Let $\circ$ be the trust-sensitive revision operator defined with respect to the partition $\Pi = \bigcup_{i=0}^{n} \Pi_i$. For each $\Pi_i \in \Pi$, let ψ_i be a formula with $mod(\psi_i) = \Pi_i$ and let $\Psi = \{\bigvee_I \psi_i \mid I \subseteq \{0, \ldots, n\}\}$. Then $\circ$ is the function $*_A$ obtained from the trust operator that maps A to the set Ψ.*

The converse is not true. One can not start with an arbitrary trust operator, and then define an equivalent partition-based revision operator.

3.3. **Trust over a Formula**

As noted previously, our motivating example has the property that agent C is trusted when they assert E, but they are not trusted when they assert $\neg E$. This is certainly plausible in the example, but it does beg a question. What do we mean when we say that an agent is trusted on a particular formula ϕ? One might be tempted to assert that trust in a formula means something of the following form:

$$A \text{ trusts } B \text{ on } \phi := K *_A (B, \phi) \models \phi. \tag{3.1}$$

However, this condition is actually too weak. Informally, suppose that we trust that a particular agent can distinguish birds from mammals. However, we do not trust that they

can distinguish different kinds of birds. If this agent reports that a particular animal is an eagle, what should we believe? It seems that we should believe the animal is a bird, but we should not have definite beliefs on the type of bird. In other words, when we trust an agent on the formula ϕ, then we should believe ϕ is true if they report something *stronger* than ϕ.

We propose the following definition.

Definition 5. *A trusts B on the formula ϕ if and only if $K *_A (B, \psi) \models \phi$ for any strengthening ψ of ϕ.*

Hence, if B tells A that $\phi \wedge \gamma$ is true, then A will always believe ϕ regardless of what γ happens to assert. This is clearly a stronger condition than (3.1), which only guarantees that A will believe ϕ if they are explicitly told ϕ.

Note that trust on ϕ does not imply trust on $\neg\phi$. As seen in the motivating example, there are natural examples where it is useful to maintain this distinction.

4. Joint Reports

In this section, we move to the case of joint reports. For our purposes, a joint report is a formula ϕ that is provided by a group of agents X. We assume that ϕ has the property that each agent in X has agreed to the content of the message. No agent would agree to a message that is inconsisent with their beliefs in their area of expertise. Hence, we need not explicitly consider the decision process used to arrive at the announcement; the fact that every agent agrees to the content is sufficient for what follows. We return to this point in the discussion.

We start by extending the definition of a trust operator.

Definition 6. *Given a set $\mathbf{A}$ of agents, a joint trust operator is a function $S : 2^{\mathbf{A}} \to 2^{sent(\mathcal{L})}$ such that for all $X, Y \subseteq \mathbf{A}$, $S(X \cup Y) = S(X) \cup S(Y)$.*

So a joint trust operator maps each set of agents to set of sentences. However, the mapping is restricted so that the set of sentences associated with each group is built up from the trust held in the members.

We also extend the notion of a trust scenario to a *joint trust scenario* by including a joint trust operator rather than a simple trust operator. Finally we extend the definition of a report to allow sources that are sets of agents, rather than single agents.

Definition 7. *Let $\langle \mathbf{A}, R, S\rangle$ be a joint trust scenario. For any belief set K, agent A, set of agents X, and formula ϕ, define:*

$$K *_A (X, \phi) = K *_A \phi[S(X)].$$

So now we are revising by the strongest weakening of ϕ over which X is considered an authority. In terms of the structure of the group of announcements, it is clear that the collection of possible *trusted sentences* for the group defines a lattice.

This characterization of trust leads directly to some nice properties. First, we can specify what it means to trust a group on ϕ in terms of the trust held in subgroups.

Proposition 2. *A trusts X on ϕ if and only if for any ψ such that $\psi \models \phi$, we have $K *_A (Y, \psi) \models \phi$ for some $Y \subseteq X$.*

Hence, a group of agents is trusted on ϕ just is case some subset of the group is trusted on each strengthening of ϕ. Moreover, since we have defined joint revision in terms of weakenings, we have the following result.

Proposition 3. *If $\phi \in (K *_A (X, \phi) \cap K *_A (Y, \phi))$, then $K *_A (X, \phi) = K *_A (Y, \phi)$.*

This result says that, if ϕ is believed when reported by X and it is believed when reported by Y, then revision by ϕ gives the same result for both groups.

We conclude this section by pointing out that the present framework does not capture all of our intuitions about joint trust. In practical examples, we would like to be able to aggregate individual trust to calculate the trust held in a group. For example, a group that includes a medical doctor and a chef should be trusted to know about the nutritional value of a particular dish. Formally, we would like to have a result of the form:

$$K *_A (X \cup Y, \phi) = Cn((K *_A (X, \phi) \cup (K *_A (Y, \phi)).$$

But in the present framework, this is not always true. The sets defined by S are not constrained at all; they need not even be closed under conjunction. In order to ensure conditions like the one above hold, we need to add some additional constraints on S. We leave the specification of these constraints and the precise characterization of suitable trust operators for future work.

5. Discussion

We have introduced a simple notion of trust over a formula, as well as an approach for dealing with reports that come from groups of agents. We have already pointed to one direction for future research: we need to give precise constraints on the trust operators in order to capture different notions of trust that are suitable for different areas of application.

Another direction for future work is to formalize the manner in which the set of reporting agents agrees on announcements. In the present framework, we simply assume that joint announcements are consistent with the expertise of each member of the reporting group. But in practice, the manner in which a group decides on announcements is important. If a group takes a democratic approach, for example, then the announcements may not reflect the expertise of a single individual. Similarly, an individual with more power or influence could sway the announcements made in a way that is not consistent with overall expertise. Fundamentally, aspects of social choice theory need to play a role in the formulation of announcements. It follows therefore that the agent receiving the announcement will need to have some knowledge of these processes in order to incorporate new information accurately.

References

[1] A. Hunter. "Belief Revision with Dishonest Reports". In: *35th Australasian Joint Conference*. 2022, pp. 397–410.

[2] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. "On The Logic of Theory Change: Partial Meet Functions for Contraction and Revision". In: *Journal of Symbolic Logic* 50.2 (1985), pp. 510–530.

[3] H. Katsuno and A. Mendelzon. "Propositional Knowledge Base Revision and Minimal Change". In: *Artificial Intelligence* 52.2 (1992), pp. 263–294.

[4] R. Booth and A. Hunter. "Trust as a Precursor to Belief Revision". In: *J. Artif. Intell. Res.* 61 (2018), pp. 699–722.

[5] F. Liu and E. Lorini. "Reasoning About Belief, Evidence and Trust in a Multi-agent Setting". In: *PRIMA 2017: Principles and Practice of Multi-Agent Systems - 20th International Conference*. 2017, pp. 71–89.

[6] J. Jiang and P. Naumov. "In Data We Trust: The Logic of Trust-Based Beliefs". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*. 2022, pp. 2683–2689.

[7] J. Singleton and R. Booth. "Who's the Expert? On Multi-source Belief Change". In: *Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR*. 2022.

[8] M. Dalal. "Investigations into a Theory of Knowledge Base Revision". In: *Proceedings of the National Conference on Artificial Intelligence (AAAI)*. 1988, pp. 475–479.