Trust-Sensitive Belief Revision

Aaron Hunter
British Columbia Institute of Technology
Burnaby, Canada
aaron_hunter@bcit.ca

Richard Booth
Mahasarakham University
Mahasarakham, Thailand
ribooth@gmail.com

Abstract

Belief revision is concerned with incorporating new information into a pre-existing set of beliefs. When the new information comes from another agent, we must first determine if that agent should be trusted. In this paper, we define trust as a pre-processing step before revision. We emphasize that trust in an agent is often restricted to a particular domain of expertise. We demonstrate that this form of trust can be captured by associating a state partition with each agent, then relativizing all reports to this partition before revising. We position the resulting family of trust-sensitive revision operators within the class of *selective revision* operators of Fermé and Hansson, and we examine its properties. In particular, we show how trust-sensitive revision is *manipulable*, in the sense that agents can sometimes have incentive to pass on misleading information. When multiple reporting agents are involved, we use a distance function over states to represent differing degrees of trust; this ensures that the most trusted reports will be believed.

1 Introduction

We consider the manner in which trust impacts the process of belief revision. Many approaches to belief revision require that all new information presented for revision must be incorporated; however, this is clearly untrue in cases where information comes from an untrusted source. In this paper, we are concerned with the manner in which an agent uses an external notion of trust in order to determine how new information should be integrated with some pre-existing set of beliefs.

Our basic approach is the following. We introduce a model where each agent only trusts other agents to be able to distinguish between certain states. We use this notion of trust as a precursor to belief revision, transforming reported information so that we only revise by the part that is trusted to be correct. This is a form of *selective revision* [Fermé and Hansson, 1999]; we establish key properties of our trust-senstive revision operator, and formally introduce the notion of manipulability. We then extend our model to a more general setting, by introducing quantitative measures that allow agents to compare degrees of trust.

2 Preliminaries

2.1 Motivation

There are different reasons that an agent may or may not be trusted. In this paper, our primary focus is on trust as a function of the perceived expertise of other agents. One agent will trust information reported by another just in case they view the reporting agent as an authority capable of drawing meaningful distinctions over a particular domain. We introduce a simple motivating example, which we revisit periodically.

Example 1 Consider an agent that visits a doctor, having difficulty breathing. The agent happens to be wearing a necklace that prominently features a jewel on a pendant. During the examination, the doctor checks the patient's throat for swelling; at the same time, the doctor sees the necklace. Following the examination, the doctor tells the patient "you have a viral infection in your throat - and by the way, you should know that the jewel in your necklace is not a diamond."

Note that the doctor provides information about two distinct domains: human health and jewelry. In practice, a patient is likely to trust the doctor's diagnosis about the viral infection. However, the patient has little reason to trust the doctor's evaluation of the necklace. We suggest that a rational agent should believe the doctor's statement about the infection, while essentially ignoring the comment on the necklace. This approach is dictated by the kind of trust that the patient has in the doctor. Our aim in this paper is to formalize this kind of domain-specific trust, and then demonstrate how this form of trust is used to inform belief revision.

2.2 Belief Revision

Belief revision refers to the process in which an agent integrates new information with some pre-existing beliefs. One of the most influential approaches to belief revision is the AGM approach. This approach is defined with respect to a finite propositional vocabulary $\mathbf{F}$. A *belief set* is a deductively closed set of formulas, representing the beliefs of an agent. A revision operator is a function that takes a belief set and a formula as input, and returns a new belief set. An AGM revision operator is a revision operator that satisfies the AGM postulates, as specified in [Alchourrón *et al.*, 1985].

A *state* is a propositional interpretation over $\mathbf{F}$, and we write $2^{\mathbf{F}}$ for the set of all states. It turns out that every AGM revision operator is characterized by a total pre-order over states. To be more precise, a *faithful assignment* is a function that maps each belief set to a total pre-order over states in which the models of the belief set are minimal. When an agent is presented with a new formula ϕ for revision, the revised belief set is given by the set of all minimal models of ϕ in the total pre-order given by the faithful assignment. We refer the reader to [Katsuno and Mendelzon, 1992] for a proof of this result, and a discussion of the implications. At present, we simply need to know that each AGM revision operator is associated with a faithful assignment.

3 A Model of Trust

3.1 Domain-Specific Trust

Assume a fixed propositional signature $\mathbf{F}$ and a set of agents $\mathbf{A}$. For each $A \in \mathbf{A}$, let $*_A$ denote an AGM revision operator. This revision operator represents an "ideal" revision, in which A has complete trust in the new information. We want to modify the way this operator is used, by adding a representation of trust with respect to each agent $B \in \mathbf{A}$.

We assume that all new information is reported by an agent, so each formula for revision can be labelled with the name of the reporting agent.[1] At this point, we are not concerned with degrees of trust or with conflicts between different sources. We start with a binary notion of trust, where A either trusts B or does not trust B with respect to a particular domain.

We encode trust by allowing each agent A to associate a partition Π^B_A over possible states with each agent B.

Definition 1 *A* state partition Π *is a collection of subsets of $2^{\mathbf{F}}$ that is collectively exhaustive and mutually exclusive. For any $s \in 2^{\mathbf{F}}$, let $\Pi(s)$ denote the element of* Π *containing s.*

If $\Pi = \{2^{\mathbf{F}}\}$, we call Π the *trivial partition* with respect to $\mathbf{F}$. If $\Pi = \{\{s\} \mid s \in 2^{\mathbf{F}}\}$, we call Π the *unit partition*.

Definition 2 *For each $A \in \mathbf{A}$ the* trust function *T_A is a function that maps each $B \in \mathbf{A}$ to a state partition Π^B_A.*

The partition Π^B_A specifies the states that A will trust B to distinguish. If $\Pi^B_A(s_1) \neq \Pi^B_A(s_2)$, then A will trust that B can distinguish between states s_1 and s_2. Conversely, if $\Pi^B_A(s_1) = \Pi^B_A(s_2)$, then A does not see B as an authority capable of distinguishing between s_1 and s_2. We clarify by returning to our motivating example.

Example 2 Let $\mathbf{A} = \{A, D, J\}$ and let $\mathbf{F} = \{sick, diam\}$. Informally: D represents a doctor, J represents a jeweler, $sick$ is true if A has an illness ,and $diam$ is true if A is wearing a diamond. Following standard shorthand notation, we represent a state s by the set of propositional symbols that are *true* in s. We specify partitions by using the | symbol to visually separate different cells. The following partitions are intuitively plausible in this example:

$$\begin{aligned} \Pi^D_A &:= \{sick, diam\}, \{sick\} | \{diam\}, \emptyset \\ \Pi^J_A &:= \{sick, diam\}, \{diam\} | \{sick\}, \emptyset \end{aligned}$$

Thus, A trusts the doctor D to distinguish between states where A is sick as opposed to states where A is not sick. However, A does not trust D to distinguish between states that are differentiated by the authenticity of a diamond.

We emphasize that a trust partition is an agent's *perception* of the expertise of others. When the doctor says that the jewel is not a diamond, they may very well be giving an assessment that they believe is correct. In this example, it is actually reasonable to believe that the doctor feels that they can tell diamond states from not diamond states; so they are not necessarily being dishonest by providing this statement. The trust partition held by A is a reflection of A's view of the doctor; it need not be correct.

3.2 Trust-Sensitive Belief Revision

In this section, we describe how an agent A combines the revision operator $*_A$ with the trust function T_A to define a new, trust-sensitive revision operator $*^B_A$. In general, $*^B_A$ will not be an AGM operator. In particular, $*^B_A$ normally will not satisfy the *Success* postulate. This is a desirable feature.

If A is given a new formula ϕ for revision, the first thing to consider is the source B and the distinctions they are trusted to make. In other words, if A does not trust B to distinguish between states s and t, then any report from B that provides evidence for s also provides evidence for t. It follows that A need not believe ϕ after revision; A interprets ϕ to be evidence for every state s that is B-indistinguishable from a model of ϕ. The next definition helps formalize this notion.

Definition 3 *Let $T_A(B) = \Pi^B_A$. For every formula ϕ, define:*

$$\Pi^B_A[\phi] = \bigcup \{\Pi^B_A(s) \mid s \models \phi\}.$$

So $\Pi^B_A[\phi]$ is the union of all cells that contain a model of ϕ. Based on the discussion above, a report of ϕ from B is construed to be evidence for each state in $\Pi^B_A[\phi]$.

Definition 4 *Let $T_A(B) = \Pi^B_A$, and let $*_A$ be an AGM revision operator for A. For any belief set K with corresponding ordering $\prec_K$ given by the underlying faithful assignment, the trust-sensitive revision $K *^B_A \phi$ is the set of formulas true in*

$$\min_{\prec_K}(\{s \mid s \in \Pi^B_A[\phi]\}).$$

So rather than taking the minimal models of ϕ, we take all minimal states among those that B can not be trusted to distinguish from the models of ϕ.

Example 3 Returning to our example, we consider a few different formulas for revision:

$$\phi_1 = sick; \qquad \phi_2 = \neg diam; \qquad \phi_3 = sick \wedge \neg diam.$$

Assume the initial belief set is given by the pre-order $\prec_K$:

$$\{diam\} \prec_K \{sick, diam\}, \emptyset \prec_K \{sick\}.$$

We have the following results for revision:

(1) $K *^D_A \phi_1 = Cn(sick \wedge diam)$.

(2) $K *^D_A \phi_2 = Cn(\neg sick \wedge diam)$.

(3) $K *^D_A \phi_3 = Cn(sick \wedge diam)$.

Result (1) shows that A believes when the doctor says that they are sick; (2) indicates the doctor is not believed on the subject of jewelry. Finally, (3) shows that an agent is able to incorporate a part of a formula, because A only incorporates the part of ϕ_3 over which the doctor is trusted.

[1] In domains involving sensing or other forms of discovery, we allow an agent A to self-report information with complete trust.

4 Formal Properties

4.1 Basic Results

We first consider extreme cases for trust-sensitive revision. Intuitively, if $T_A(B)$ is the trivial partition, then A does not trust B to be able to distinguish between any states. Hence, A should not incorporate any information obtained from B.

Proposition 1 *If $T_A(B)$ is the trivial partition, then $K *_A^B \phi = K$ for all K and ϕ.*

The other extreme situation occurs when $T_A(B)$ is the unit partition, consisting of all singleton sets. In this case, A trusts B to distinguish between every possible pair of states.

Proposition 2 *If $T_A(B)$ is the unit partition, then $*_A^B = *_A$.*

Hence, if B is universally trusted, then trust-sensitive revision is just normal AGM revision.

Partitions are partially ordered by *refinement*. We say that Π_1 is a refinement of Π_2 just in case, for each $S_1 \in \Pi_1$, there exists $S_2 \in \Pi_2$ such that $S_1 \subseteq S_2$. For trust partitions, refinement has a natural interpretation as "breadth of trust."

Proposition 3 *For any formula ϕ, if Π_A^B is a refinement of Π_A^C, then $K *_A^C \phi \subseteq K *_A^B \phi$.*

So if B is trusted over a greater range of states, then receiving ϕ from B yields a larger belief set than receiving ϕ from C.

4.2 Trust-Sensitive Revision as Selective Revision

Trust-sensitive revision is a specialized version of *selective revision* [Fermé and Hansson, 1999]. An operator $\circ$ is a selective revision operator if there exists an AGM revision operator $*$ and a *transformation function* f taking formulas to formulas such that, for all belief sets K and formulas ϕ,

$$K \circ \phi = K * f(\phi).$$

The operator $*_A^B$ clearly falls under this scheme. It's the particular instance obtained by allowing f to be defined via the state partition Π_A^B with $f(\phi) = \phi_A^B$ such that the models of ϕ_A^B are precisely the models in $\Pi_A^B[\phi]$.

Fermé and Hansson enumerate several properties for the function f and prove correspondence results between properties of f and postulates for $\circ$. These results allow us to give a list of sound postulates for trust-sensitive revision. The relevant properties of f from [Fermé and Hansson, 1999] are as follows[2] ($\vdash$ and $\equiv$ denote classical logical consequence and equivalence respectively):

- $f(\bot) \equiv \bot$ (falsity preservation)
- $\phi \vdash f(\phi)$ (implication)
- $f(f(\phi)) \equiv f(\phi)$ (idempotence)
- If $\phi \equiv \psi$ then $f(\phi) \equiv f(\psi)$ (extensionality)
- $f(\phi \vee \psi) \equiv f(\phi) \vee f(\psi)$ (disjunctive distribution)

The above properties are familiar from topology. They essentially express that f is a *Kuratowski closure operator* on the space of subsets of states [Kuratowski, 1958]. We thus make the following definition.

Definition 5 *A function f taking formulas to formulas satisfying the above five properties will be called a* Kuratowski transformation function.

The next result is proved in [Fermé and Hansson, 1999].

Proposition 4 ([Fermé and Hansson, 1999]) *Let $*$ be an AGM revision operator and f be a Kuratowski transformation function. Then $\circ$ derived from $*$ and f satisfies all the following postulates.*

- *$K \circ \phi = Cn(K \circ \phi)$ (Closure)*
- *There is a formula ψ such that $K \circ \phi \vdash \psi$, $\phi \vdash \psi$ and $K \circ \phi = K \circ \psi$ (Proxy success)*
- *$K \circ \phi \subseteq Cn(K \cup \{\phi\})$ (Inclusion)*
- *$K \circ \phi$ is consistent iff ϕ is consistent (Consistency)*
- *If $\phi \equiv \psi$ then $K \circ \phi = K \circ \psi$ (Extensionality)*
- *If $K \not\subseteq K \circ \phi$ then $K \cup (K \circ \phi) \vdash \bot$ (Consistent expansion)*
- *$(K \circ \phi) \cap (K \circ \psi) \subseteq K \circ (\phi \vee \psi)$ (Disjunctive overlap)*
- *If $K \circ (\phi \vee \psi) \not\vdash \neg\phi$ then $K \circ (\phi \vee \psi) \subseteq K \circ \phi$ (Disjunctive inclusion)*
- *$K \circ (\phi \vee \psi)$ is equal to one of $K \circ \phi$, $K \circ \psi$ or $(K \circ \phi) \cap (K \circ \psi)$ (Disjunctive factoring)*

The above postulates are mostly familiar from the literature on belief change. The *Success* postulate is replaced by the weaker *Proxy success*.

What, then, are the properties of trust-sensitive revision? We can immediately state the following.

Proposition 5 *Let f be defined via a state partition Π. Then f is a Kuratowski transformation function. Thus every trust-sensitive revision operator satisfies all the postulates listed in Proposition 4.*

Trust-sensitive revision also satisfies a new postulate:

Proposition 6 *Every trust-sensitive revision operator satisfies the following postulate:*

- *There exist $\lambda_1, \ldots, \lambda_m$ such that* (i) $(K \circ \lambda_i) \cup (K \circ \lambda_j) \vdash \bot$ *for $i \neq j$, and* (ii) *for all ϕ there exists a set $X \subseteq \{1, \ldots, m\}$ such that $K \circ \phi = \bigcap_{i \in X} K \circ \lambda_i$ (Disjoint outcome basis)*

Disjoint outcome basis says there is some finite basic set of mutually inconsistent revision outcomes $K \circ \lambda_1, \ldots, K \circ \lambda_m$ such that *every* revision outcome can be expressed as an intersection of them. In fact, roughly speaking, we may take the λ_i to be the formulas that are closed under f, i.e., such that $f(\lambda_i) \equiv \lambda_i$. It can be shown that this postulate does not hold in general for selective revision operators defined via Kuratowski transformation functions.

As we have seen, *Success* does not hold in general for trust-sensitive revision. Even the following weaker version (which is also implied by another of the basic AGM revision postulates, namely *Vacuity*) fails [Hansson, 1997]:

- If $K \not\vdash \neg\phi$ then $K \circ \phi \vdash \phi$ (Weak Success)

[2] The first property here does not actually appear in their paper.

As an extreme counterexample to $*_A^B$ failing *Weak Success*, just take $T_A(B)$ to be the trivial partition so that $K *_A^B \phi$ does not differ from K for *any* choice of ϕ. The failure of this rule is a departure from Fermé and Hansson, who hold onto it by always assuming the transformation function satisfies $f(\phi) \equiv \phi$ whenever $K \nvdash \neg\phi$. We argue this assumption makes little sense in our setting: why should we accept information from an untrusted source just because it happens to be consistent with our current beliefs? Is there an even weaker variant of *Weak Success* that holds for trust-sensitive revision? As a first idea, one might suggest the following:

- If $K \nvdash \neg\phi$ and $K \circ \neg\phi \vdash \neg\phi$ then $K \circ \phi \vdash \phi$
(Very Weak Success)

This rule roughly says that, if we trust B when it tells us ϕ is false, then we should trust B when it tells us ϕ is true. The decision on whether to accept B's information about ϕ is judged on B's perceived ability to answer the yes-or-no question of whether ϕ holds. This intuition is perhaps clearer in the following equivalent formulation.

- If $K \nvdash \neg\phi$ and $K \nvdash \phi$ then $[K \circ \phi \vdash \phi$ iff $K \circ \neg\phi \vdash \neg\phi]$

Trust-sensitive revision fails this postulate too, in general:

Proposition 7 *There exists an AGM revision operator $*$ and a state partition Π such that $\circ$ defined via $*$ and Π does not satisfy* Very Weak Success.

This result follows from the following counterexample, which is given further motivation below.

Example 4 Suppose $\mathbf{F} = \{p, q\}$, let $K = Cn(q)$, and let $\prec_K$ be the total pre-order where the models of K are minimal and everything else is maximal. Suppose that the trust partition is $\Pi = \{p, q\} \mid \{q\}, \{p\} \mid \emptyset$. Then we have: $K \nvdash \neg p$, $K \circ \neg p = Cn(\neg p \wedge q)$, so $K \circ \neg p \vdash \neg p$. However, $K \circ p = Cn(q)$, so $K \circ p \nvdash p$.

How disappointed should we be that *Very Weak Success* fails? We argue that we need not be disappointed at all. The partition Π in the example has the property that it can only distinguish these cases: both p and q are true, neither is true, exactly one is true. Thus, if we initially believe q, then no report can ever make us believe p and $\neg q$. There is an asymmetry here: we will trust a report that p is false, but we will not trust a report that it is true.

We can frame the counterexample as the *Dead Battery Problem*. Suppose a lamp requires two batteries to make a bright light; it is dim with one good battery, and it is off with zero. Now suppose you contribute one battery and your adversary contributes the other. If your adversary tests the lamp in private and tells you that their battery works, you will not trust this conclusion. From your perspective, they may have seen the dim light and jumped to the conclusion that their battery is working. So while you would accept a report that their battery is dead, you will not accept a report that it works.

Although *Very Weak Success* does not hold, trust-sensitive revision *does* manage to capture an even weaker variant of *Success*.

Proposition 8 *Every trust-sensitive revision operator satisfies the following postulate:*

- *If $K \nvdash \neg\phi$ and $K \circ \neg\phi \vdash \neg\phi$ then there exists a consistent formula ψ such that $\psi \vdash \phi$ and $K \circ \psi \vdash \phi$. (Feeble Success)*

Feeble Success relaxes *Very Weak Success* by saying that if we trust B when it tells us ϕ is false (and we didn't already believe $\neg\phi$) then B can also bring us to believe ϕ by telling us ϕ *perhaps in conjunction with some other "extra" evidence*. For instance in the Dead Battery Problem (Example 4), although you do not accept the report of your adversary when they tell you their battery is working ($K \circ p \nvdash p$), you *would* come to believe it is working if instead they reported that *both* batteries are working ($K \circ (p \wedge q) = Cn(p \wedge q) \vdash p$).

The proof that trust-sensitive revision satisfies *Feeble Success* makes use of the fact that f defined via a state partition satisfies the following property, which in topological terms essentially says that the complement of a closed set of states is itelf closed, i.e., every open set is closed.

- $f(\neg f(\phi)) \equiv \neg f(\phi)$

Kuratowski transformation functions do not satisfy this in general, and indeed *Feeble Success* is a property not shared by every selective revision operator defined via a Kuratowski function.

Proposition 9 *There exists an AGM revision operator $*$ and a Kuratowski transformation function f such that $\circ$ defined via $*$ and f does not satisfy* Feeble Success.

This is proved from the following counterexample.

Example 5 Suppose $\mathbf{F} = \{p\}$ and let f be specified by setting $f(\bot) \equiv \bot$, $f(p) \equiv \top$, $f(\neg p) \equiv \neg p$ and $f(\top) \equiv \top$. (Every other formula in this language is equivalent to precisely one of $\bot, p, \neg p, \top$, so these four values completely specify f by appeal to *equivalence*.) One can check that f forms a Kuratowski transformation function. Assume $K = Cn(\top)$, so $K \circ \phi = K * f(\phi) = Cn(f(\phi))$ for all ϕ and any AGM revision operator $*$. Looking at f we see there is no consistent formula ψ such that $f(\psi) \vdash p$ and so there is no ψ such that $K \circ \psi \vdash p$. The consequent of *Feeble Success* (plugging in $\phi = p$) therefore cannot hold. But we have $K \nvdash \neg p$ and $K \circ \neg p = Cn(\neg p) \vdash \neg p$, thus the antecedent holds.

4.3 Manipulability

Next we consider a concept that has been extensively studied in areas such as voting theory, preference aggregation and belief merging [Chopra *et al.*, 2006; Everaere *et al.*, 2007; Gibbard, 1973; Satterthwaite, 1975], but that has not yet received attention in the belief change literature, namely *manipulability*. Let us assume that agent B, in passing information information ϕ to A, does so with the communicative *goal* of bringing about in A a belief in ϕ. Under this assumption, we can ask: *does B have any incentive to pass on any formula other than ϕ?* That is, is it possible that A will *not* believe ϕ if given ϕ directly, but *will* believe it if given some other formula ψ? The following postulate expresses that A can never be manipulated in this way.

- If $K \circ \phi \nvdash \phi$ and ψ is consistent then $K \circ \psi \nvdash \phi$
(Non-manipulability)

We remark that a slight variant of *Non-manipulability* has appeared in the literature (but with a different motivation) under the name *Regularity* [Hansson *et al.*, 2001].

Is trust-sensitive revision non-manipulable, i.e., does $\circ$ defined from an AGM revision operator and a state partition satisfy the above postulate? For our two extreme cases the answer is yes: If Π is the unit partition then $\circ$ satisfies *Success*, so of course B can then do no better than just telling ϕ to A to achieve its goal, while if Π is the trivial partition then B can say nothing at all to change A's beliefs so in both cases the postulate is trivially satisfied. However, in general the answer is no, which can be seen from the following.

Proposition 10 *If a revision operator* $\circ$ *satisfies both* Feeble Success *and* Non-Manipulability *then it satisfies* Very Weak Success.

Since we have already seen that trust-sensitive revision satisfies *Feeble Success* but *not*, in general *Very Weak Success* (Propositions 7 and 8), we can immediately state:

Proposition 11 *There exists an AGM revision operator* $*$ *and a state partition* Π *such that* $\circ$ *defined via* $*$ *and* Π *fails* Non-manipulability.

So trust-sensitive revision is *manipulable*. However, this is neither surprising nor undesirable. We already showed that it is not manipulable in the extreme cases. But informally, the notion of trust means that one agent is willing to believe things said by another. If we trust an agent to be able to draw certain distinctions, then we also accept the consequences of those distinctions when incorporating their reports.

5 Trust Pseudometrics

5.1 Measuring Trust

Thus far, we have assumed that an agent is either trusted to distinguish certain states, or they are not. This is not sufficient if we consider situations where several different sources may provide conflicting information. In such cases, we need to determine which information source is the most trusted.

In order to capture different levels of trust, we introduce a measure of the distance between states. Each agent A will associate a distance function d_B over states with each agent B. If $d_B(s,t) = 0$, then B can not be trusted to distinguish between the states s and t. If $d_B(s,t)$ is large, then A has a high level of trust in B's ability to distinguish between s and t. We will require that the distance function is a *pseudo-ultrametric* on the state space, which means that it satisfies the following properties for all x, y, z:

1. $d(x,x) = 0$
2. $d(x,y) = d(y,x)$
3. $d(x,z) \leq \max\{d(x,y), d(y,z)\}$

A pseudo-ultrametric differs from an *ultrametric* in that $d(x,y) = 0$ does not imply $x = y$; this would be undesirable since we use the distance 0 to represent indistinguishability rather than identity. The third property is the *ultrametric inequality*, which is a strengthening of the triangle inequality.

Definition 6 *If* $A \in \mathbf{A}$, *an* ultrametric trust function $\mathcal{T}_A$ *is a function mapping each* $B \in \mathbf{A}$ *to a pseudo-ultrametric* d_B *on* $2^{\mathbf{F}}$.

The pair $(2^{\mathbf{F}}, \mathcal{T}_A)$ is called a *trust space*. We associate a sequence of state partitions with each trust space.

Definition 7 *Let* $(2^{\mathbf{F}}, \mathcal{T}_A)$ *be a trust space, let* $B \in \mathbf{A}$, *and let* i *be a number. For each state* s, *define:* $\Pi_A^B(i)(s) = \{t \mid d_B(s,t) \leq i\}$.

The following proposition is a known result in the theory of *metric spaces*, restated in our setting. The proof requires the ultrametric inequality.

Proposition 12 *For any number* i, *the collection of sets* $\{\Pi_A^B(i)(s) \mid s \in 2^{\mathbf{F}}\}$ *is a state partition. Moreover, if* $i < j$, *then* $\Pi_A^B(i)$ *is a refinement of* $\Pi_A^B(j)$.

The cells of the partition $\Pi_A^B(i)$ consist of all states that are separated by a distance of no more than i. Hence, a pseudometric trust space defines a sequence of partitions for each agent that gets coarser as we increase the index i. Since we can use Definition 4 to define a trust-sensitive revision operator from a state partition, we can now define a trust-sensitive revision operator for any fixed distance i between states. The simplest such operator is the following.

Definition 8 *Let* $(2^{\mathbf{F}}, \mathcal{T}_A)$ *be a trust space. The* trust-sensitive revision operator *for* A *with respect to* B *is the trust-sensitive revision operator given by* $\Pi_A^B(0)$.

In other words, given an ultrametric, the simplest revision operator is obtained by saying states are indistinguishable just in case the distance between them is 0. However, as i increases, we get different partitions that require greater trust in B to distinguish between states. In the next section, we use this sequence of partitions to resolve conflicting reports between agents that are trusted to differing degrees.

Example 6 We give two ultrametrics that produce the same basic trust partition. Assume two doctors: a general practitioner D and a specialist S. The vocabulary includes two symbols: ear and $skin$. Informally, ear is true if the patient has an ear infection, and $skin$ is true if the patient has skin cancer. An ear infection can be diagnosed by any doctor, whereas skin cancer is typically diagnosed by a specialist. To capture these facts, we define two pseudo-ultrametrics d_D and d_S over the states labelled as follows:

$$s_1 = \{ear, skin\}; \quad s_2 = \{ear\}; \quad s_3 = \{skin\}; \quad s_4 = \emptyset.$$

	s_1, s_2	s_1, s_3	s_1, s_4	s_2, s_3	s_2, s_4	s_3, s_4
d_D	1	2	2	2	2	1
d_S	2	2	2	2	2	2

Note that S is more trusted than D to distinguish states related to skin cancer. As such, S should be believed in the case of conflicting reports from D and S with respect to such states.

5.2 Multiple Reports

Formally, define a *report* to be a pair (ϕ, B) where ϕ is a formula and $B \in \mathbf{A}$. For simplicity in this section we assume ϕ is consistent and also that, given a finite set of reports $\Phi = \{(\phi_i, B_i) \mid i < n\}$ incoming to A, we have $B_i \neq B_j$ for $i \neq j$. We are interested in using ultrametric trust functions to address the situation where an agent simultaneously receives a set of reports from different agents.

Proposition 13 *Let $\{A\} \cup \mathbf{B} \subseteq \mathbf{A}$, and let $\Phi = \{(\phi_i, B_i) \mid i < n\}$ be a finite set of reports. There exists a number m such that $\bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i]) \neq \emptyset$.*

Hence, for any set of reports, we can get a non-empty intersection if we take a sufficiently coarse state partition. In many cases this partition will be non-trival. Using this proposition, we define multiple report revision as follows.

Definition 9 *Let $(2^{\mathbf{F}}, \mathcal{T}_A)$ be a trust space, let $\Phi = \{(\phi_i, B_i) \mid i < n\}$ be a finite set of reports, and let m be the least number such that $\bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i]) \neq \emptyset$. Given K, $*$ and $\prec_K$, the trust-sensitive revision $K *_A^{\mathbf{B}} \Phi$ is the set of formulas true in*

$$\min_{\prec_K}(\{s \mid s \in \bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i]\}).$$

Hence, trust-sensitive revision involves finding the finest partition that provides a meaningful combination of the reports, and then revising with the corresponding state partition.

Example 7 Continuing Example 6, suppose $\Phi = \{(ear \wedge skin, D), (\neg skin, S)\}$. Then the least m such that $\Pi_A^D(m)[ear \wedge skin] \cap \Pi_A^S(m)[\neg skin] \neq \emptyset$ is equal to 1, for which we have $\Pi_A^D(1)[ear \wedge skin] \cap \Pi_A^S(1)[\neg skin] = \{s_1, s_2\} \cap \{s_2, s_4\} = \{s_2\}$. Thus we make an AGM revision by $ear \wedge \neg skin$, i.e., we ignore the part of D's report concerning $skin$ before consistently conjoining the 2 reports.

This example demonstrates how ultrametrics can be used to resolve conflicts by appealing to strength of trust.

The sub-procedure of incrementally weakening the formulas ϕ_i until consistency is reached is highly reminiscent of the *belief negotiation* approach to belief merging [Booth, 2006]. Indeed the specific procedure used here can be viewed as a variant of one of the negotiation protocols defined there, which in turn yields a variant of the $\Delta^{\max}$ merging operator of [Konieczny and Pino Pérez, 2002]. This connection allows us to describe the set $\bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i])$ in Definition 9 directly in terms of the pseudo-ultrametrics d_{B_i}. First some notation. Given the finite set of reports $\Phi = \{(\phi_i, B_i) \mid i < n\}$, for each $i < n$ and $s \in 2^{\mathbf{F}}$ define $R_i(s) = \min\{d_{B_i}(s,t) \mid t \text{ is a model of } \phi_i\}$, i.e., $R_i(s)$ represents the degree to which B_i can be trusted to distinguish s from a model of ϕ_i. Then we have the following.

Proposition 14 *Let $\Phi = \{(\phi_i, B_i) \mid i < n\}$ be a finite set of reports, and let m be the least number such that $\bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i]) \neq \emptyset$. Then $\bigcap_{i<n}(\Pi_A^{Bi}(m)[\phi_i]) = \{s \mid \max_{i<n} R_i(s)$ is minimal$\}$.*

6 Discussion

6.1 Related Work

The relationship between trust and belief change is explored in [Lorini *et al.*, 2014], through a dynamic epistemic logic(DEL) called DL-BT in which one can express that an agent A trusts another agent B to be able to determine the truth of a formula ϕ with strength α. In DL-BT, the emphasis is on iterated belief change and update policies for orderings. By contrast, we explicitly build on the AGM approach, with an emphasis on single-shot revision. Since DL-BT is a modal logic, there are many superficial differences from our approach both in terms of syntax and semantics. More importantly, DL-BT differs from our approach in that it defines trust with respect to an agent's ability to determine the truth or falsity of a *formula*, whereas we define trust with respect to *distinguishable states*. Indeed, we have seen that trust-sensitive revision does not even satisfy *Very Weak Success*; an agent may be trusted on ϕ without being trusted on $\neg\phi$. It is possible to define trust-sensitive revision based on state partitions in a DEL setting, and the resulting logic differs from DL-BT in a non-trivial manner. We leave the exploration of our approach in a DEL setting for future work.

Another epistemic logic approach is [Rodenhäuser, 2014]. In that work, a commonality with our approach is that each agent A is assumed to assign a kind of *indicator* to every other agent B to denote the level of trust in B. But rather than use a state partition or pseudo-ultrametric as we do, the indicator takes the form of a particular *plausibility revision policy* to reflect the strength with which A should incorporate information from B into its belief state. However the kind of trust modeled is more like *credibility* or *reliability* than expertise. This last comment also applies to the *credibility-limited revision* operators studied in [Hansson *et al.*, 2001].

Additional work on trust has explored reputation systems [Huynh *et al.*, 2006], task allocation [Ramchurn *et al.*, 2009], and deception [Salehi-Abari and White, 2009]. One notable approach with an emphasis on knowledge representation is [Wang and Singh, 2007], in which trust is built based on evidence; this could be used as a precursor step to build a trust metric in our framework. Different levels of trust are treated in [Krukow and Nielsen, 2007], where a lattice structure is used to represent various levels of trust strength. However, the emphasis is on the representation of trust in *an agent* as opposed to trust in an agent with respect to a domain.

6.2 Conclusion

We have developed an approach to trust-sensitive belief revision in which trust is captured by state partitions. Trust is handled as a precursor to belief change; the input formula is relativized to the trust partition prior to revision. The result is a form of selective revision that satisfies many known postulates for belief change, but not others. In particular, it fails to satisfy all but the most extreme weakenings of the success postulate. Additionally it turns out to falsify our new *Non-manipulability* property. We briefly address the manner in which conflicting reports can be addressed by introducing an ultrametric on states that can capture a relative degree of trust.

There are many directions for future work. First, we would like to explicitly characterize trust-sensitive revision by precisely specifying the underlying class of Kuratowski transformations. There is also the question of iteration. In principle, we can relativize any mapping on orderings based on a state partition; however, further investigation is required to determine the properties of this approach to iterated revision, because there is flexibility in handling indistinguishable states at different levels of an underlying pre-order. We would also like to address the issue of *deception* in greater detail; this has many applications in security and networked communication.

Acknowledgements

Richard Booth's contribution was written while visiting the Individual and Collective Reasoning group at the University of Luxembourg, whose support is gratefully acknowledged. Thanks are due to Emil Weydert for pointing us to the links with topology.

References

[Alchourrón *et al.*, 1985] Carlos Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[Booth, 2006] Richard Booth. Social contraction and belief negotiation. *Information Fusion*, 7(1):19–34, 2006.

[Chopra *et al.*, 2006] Samir Chopra, Aditya Ghose, and Thomas Meyer. Social choice theory, belief merging, and strategy-proofness. *Information Fusion*, 7(1):61–79, 2006.

[Everaere *et al.*, 2007] Patricia Everaere, Sébastien Konieczny, and Pierre Marquis. The strategy-proofness landscape of merging. *Journal of Artificial Intelligence Research*, 28:49–105, 2007.

[Fermé and Hansson, 1999] Eduardo Fermé and Sven Ove Hansson. Selective revision. *Studia Logica*, 63(3):331–342, 1999.

[Gibbard, 1973] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

[Hansson *et al.*, 2001] Sven Ove Hansson, Eduardo Fermé, John Cantwell, and Marcelo Falappa. Credibility-limited revision. *The Journal of Symbolic Logic*, 66(04):1581–1596, 2001.

[Hansson, 1997] Sven Ove Hansson. Semi-revision. *Journal of Applied Non-Classical Logics*, 7(1-2):151–175, 1997.

[Huynh *et al.*, 2006] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[Katsuno and Mendelzon, 1992] Hirofumi Katsuno and Alberto Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(2):263–294, 1992.

[Konieczny and Pino Pérez, 2002] Sébastien Konieczny and Ramón Pino Pérez. Merging information under constraints: a logical framework. *Journal of Logic and Computation*, 12(5):773–808, 2002.

[Krukow and Nielsen, 2007] Karl Krukow and Mogens Nielsen. Trust structures. *International Journal of Information Security*, 6(2-3):153–181, 2007.

[Kuratowski, 1958] Kazimierz Kuratowski. *Topology Volume 1*. Academic Press, 1958.

[Lorini *et al.*, 2014] Emiliano Lorini, Guifei Jiang, and Laurent Perrussel. Trust-based belief change. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, pages 549–554, 2014.

[Ramchurn *et al.*, 2009] Sarvapali D. Ramchurn, Claudio Mezzetti, Andrera Giovannucci, Juan A. Rodriguez-Aguilar, Rajdeep K. Dash, and Nicholas R. Jennings. Trust-based mechanisms for robust and efficient task allocation in the presence of execution uncertainty. *JAIR*, 35:119–159, 2009.

[Rodenhäuser, 2014] Ben Rodenhäuser. *A Matter of Trust: Dynamic Attitudes in Epistemic Logic*. PhD thesis, University of Amsterdam, 2014.

[Salehi-Abari and White, 2009] Amirali Salehi-Abari and Tony White. Towards con-resistant trust models for distributed agent systems. In *IJCAI*, pages 272–277, 2009.

[Satterthwaite, 1975] Mark Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

[Wang and Singh, 2007] Yonghong Wang and Munindar P. Singh. Formal trust model for multiagent systems. In *IJCAI*, pages 1551–1556, 2007.