Journal of Artificial Intelligence Research 61 (2018) 699-722 Submitted 03/17; published 03/18

Trust as a Precursor to Belief Revision

Richard Booth BOOTHR2@CARDIFF.AC.UK
Cardiff University
Cardiff, UK

Aaron Hunter AARON_HUNTER@BCIT.CA
British Columbia Institute of Technology
Burnaby, Canada

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 prove a representation result that characterizes the class of trust-sensitive revision operators in terms of a set of postulates. We also show that trust-sensitive revision is *manipulable*, in the sense that agents can sometimes have incentive to pass on misleading information.

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 & Hansson, 1999); we prove that our trust-sensitive revision operators can be characterized by a natural set of rationality postulates. We establish key properties of our trust-senstive revision operator, and formally introduce the notion of manipulability. This paper is an extended version of our previous work (Hunter & Booth, 2015).

2. Preliminaries

We begin with our motivation and some brief background on belief revision.

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 a patient that visits a doctor, having difficulty breathing. The patient 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. As such, we suggest that a rational agent should believe the doctor's statement about the infection, while essentially ignoring the comment on the necklace. 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. AGM revision is defined with respect to a propositional vocabulary $\mathbf{F}$; in this paper, we assume that $\mathbf{F}$ is finite. A *belief set* is a deductively closed set of propositional 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 (Alchourrón, Gärdenfors, & Makinson, 1985).

A *state* is a propositional interpretation over $\mathbf{F}$, and we write $2^{\mathbf{F}}$ for the set of all states. Following standard conventions, we let $\top$ stand for a formula that is true in all states and we let $\bot$ stand for a formula that is not true in any state. 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

We now build towards our model of trust-sensitive belief revision.

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 provided 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 $\Pi_{B,K}$ over possible states with each agent B and each belief set K.

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* $B \in \mathbf{A}$ *and each belief set* K *to a state partition* $\Pi_{B,K}$*.*

The partition $\Pi_{B,K}$ specifies the states that A will trust B to distinguish, given A's current belief set is K. If $\Pi_{B,K}(s_1) \neq \Pi_{B,K}(s_2)$, then A will trust that B can distinguish between states s_1 and s_2. Conversely, if $\Pi_{B,K}(s_1) = \Pi_{B,K}(s_2)$, 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:

$$\begin{aligned} \Pi_{D,K} &:= \{sick, diam\}, \{sick\} | \{diam\}, \emptyset \\ \Pi_{J,K} &:= \{sick, diam\}, \{diam\} | \{sick\}, \emptyset \end{aligned}$$

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. The partitions can be visualized as in Figure 1.

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

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

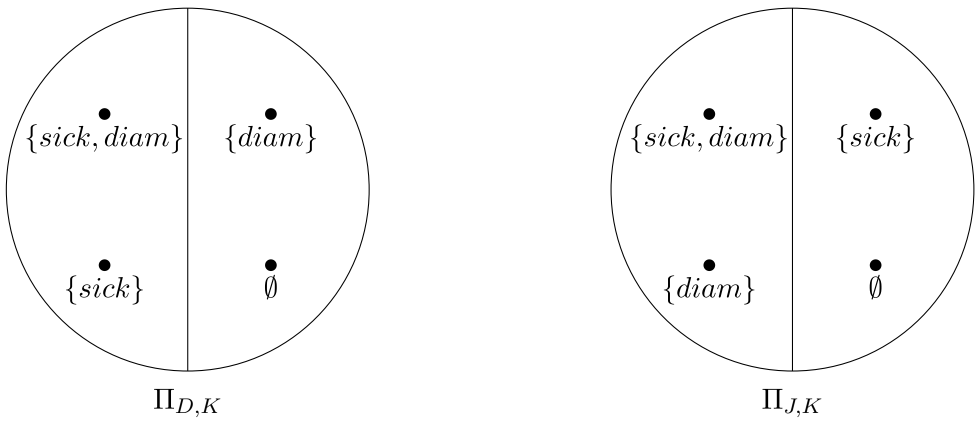

Figure 1: Visualizing Trust Partitions

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.

We remark that our notion of indistinguishability here is transitive: if B cannot distinguish between the pair s_1, s_2 and B also cannot distinguish between the pair s_2, s_3, then it follows that B cannot distinguish between s_1 and s_3. This property might not be desirable if we think of indistinguishability in terms of B's perception or ability to discern minor differences between states. But recall that our state partitions are actually defined to capture some agent A's view of B's expertise. If A does not trust B to distinguish s_1 and s_2, this means that A views B as incapable of telling the difference between s_1 and s_2. From A's perspective, any time B says the state is s_1, it is equally likely that the actual state is s_2. From this perspective, the transitive property of indistinguishability is plausible.

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 family of trust-sensitive revision operators $*^B_A$, one for each $B \in \mathbf{A}$. 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, that we discuss later in the present paper.

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* Π *be a state partition. For every formula* ϕ*, define:*

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

So $\Pi[\phi]$ is the union of all cells that contain a model of ϕ. Based on the discussion above, a report of ϕ from B is interpreted as evidence for each state in $\Pi_{B,K}[\phi]$.

Definition 4 *Let $T_A(K, B) = \Pi_{B,K}$, 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 $K *_A^B \phi$ is the set of formulas true in*

$$\min_{\prec_K}(\{s \mid s \in \Pi_{B,K}[\phi]\}).$$

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

This notion can be formulated syntactically as well. Since $\mathbf{F}$ is finite, each state s is defined by a unique, maximal conjunction over literals.

Definition 5 *For any state s, let $prop(s)$ denote the unique, maximal conjunction of literals true in s.*

This definition can be extended for a cell in a state partition.

Definition 6 *Let Π be a state partition. For any state s,*

$$prop(\Pi(s)) = \bigvee\{prop(s') \mid s' \in \Pi(s)\}.$$

Note that $prop(\Pi(s))$ is a well-defined formula in disjunctive normal form, due to the finiteness of $\mathbf{F}$. Intuitively, $prop(\Pi(s))$ is the formula that defines the set of states $\Pi(s)$. In the case of a trust partition $\Pi_{B,K}$, we can use this idea to define the *trust expansion* of a formula.

Definition 7 *The* trust expansion *of ϕ for A with respect to B is the formula $\phi^A_K(B) := \bigvee\{prop(\Pi_{B,K}(s)) \mid s \models \phi\}$.*

Note that $\phi^A_K(B)$ is true in all states that are consistent with ϕ with respect to distinctions that A trusts B to make. Trust-sensitive revision can equivalently be defined by translating ϕ to $\phi^A_K(B)$, then performing AGM revision.

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 $\prec_K$:

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

Figure 2 shows the partition cells that contain the models of each of these formulas, by shading the complete cell that will be used for revision. As such, we have the following results for revision:

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

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

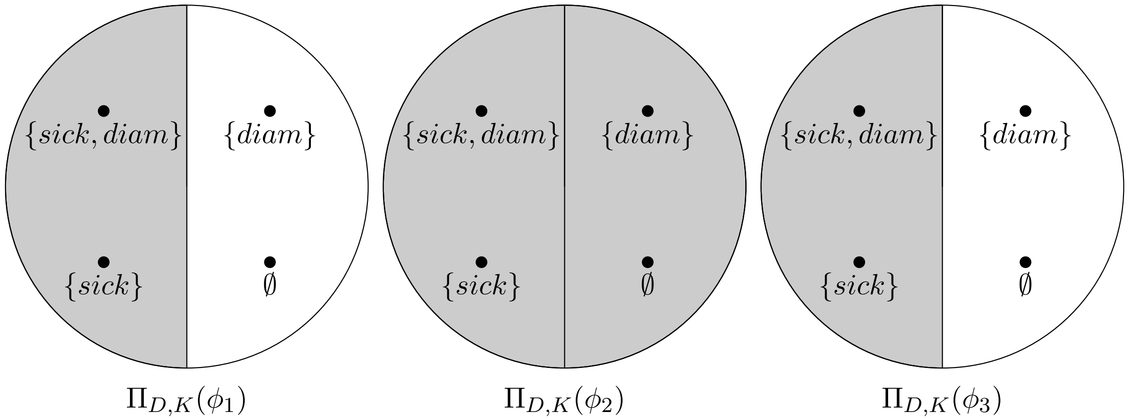

Figure 2: Partition cells containing the models of the input formulas

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

Result (1) shows that A believes when the doctor says that they are sick; (2) indicates that 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.

4. Formal Properties

In this section we consider the properties of our trust-sensitive revision operators.

4.1 Basic Results

We first consider extreme cases for trust-sensitive revision. Intuitively, if $T_A(K, B)$ 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. In the proof of the following proposition, and throughout the rest of the paper, $|X|$ is used to denote the set of models of a set X of formulas.

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

Proof Suppose $T_A(K, B) = \{2^{\mathbf{F}}\}$. Then $K *_A^B \phi$ is the set of formulas true in $\min_{\prec_K}(\{s \mid s \in 2^{\mathbf{F}}\})$. The minimal elements of $\prec_K$ are exactly the models of K, for any faithful assignment. $\square$

The other extreme situation occurs when $T_A(K, B)$ 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 $T_A(K, B)$ is the unit partition, then $*_A^B = *_A$.*

Proof Suppose $T_A(K, B) = \{\{s\} \mid s \in 2^{\mathbf{F}}\}$. Then $T_A(K, B)[\phi] = |\phi|$, for any formula ϕ. It follows immediately that $*_A^B = *_A$. $\square$

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 $S_1 \subseteq S_2$. We also say that Π_1 is *finer* than Π_2. For trust partitions, refinement has a natural interpretation as "breadth of trust." If the partition corresponding to B is finer than that corresponding to C, it means that B is trusted more broadly than C. To be more precise, it means that B is trusted to distinguish between all of the states that C can distinguish, and possibly more. Intuitively, if B is trusted more broadly that C, it follows that a report from B should give A more information.

4.2 Trust-Sensitive Revision as Selective Revision

Trust-sensitive revision $*_A^B$ is a specialized version of *selective revision* (Fermé & Hansson, 1999). An operator $\circ$ is a selective revision operator if there exists an AGM revision operator $*$ and, for each possible belief set K, a *transformation function* f_K taking formulas to formulas such that, for all belief sets K and formulas ϕ,

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

The operator $*_A^B$ clearly falls under this scheme. It's the particular instance obtained by allowing f_K to be defined via the state partition $\Pi_{B,K}$ with $f_K(\phi) = \phi_K^A(B)$, i.e., $f_K(\phi)$ is just the trust expansion of ϕ for A with respect to B (see Definition 7). For the remainder of this section, we use this characterization as selective revision to explore a variety of postulates that hold for trust-senstive revision. These postulates will be used to prove a representation result in §5.

Fermé and Hansson enumerate several properties for the function f_K and prove correspondence results between properties of f_K and postulates for $\circ$. These results allow us to give a list of sound postulates for trust-sensitive revision. The relevant properties of f_K 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, 1966). We thus make the following definition.

2. The first property here is not actually listed by Fermé and Hannson (1999).

Definition 8 *A function f taking formulas to formulas satisfying the above five properties will be called a* Kuratowski transformation function*.*

Another significant property of Kuratowski transformation functions (derivable from *extensionality* and *disjunctive distribution*), is as follows:

- $\phi \vdash \psi$ implies $f(\phi) \vdash f(\psi)$ (monotonicity)

The next result is proved by Fermé and Hannson (1999).[3]

Proposition 3 (Fermé & Hansson, 1999) *Let $*$ be an AGM revision operator and f_K be a Kuratowski transformation function for each K. Then $\circ$ derived from $*$ and f satisfies all the following postulates.*

- $K \circ \phi = Cn(K \circ \phi)$ *(Closure)*
- $K \circ \top = K$ *(Identity)*
- *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*.

Another two postulates that are sound for any selective revision derived from Kuratowski transformation functions are as follows:

- If $K \circ (\phi \vee \psi) \vdash \neg\phi$ then $K \circ (\phi \vee \psi \vee \theta) \vdash \neg\phi$ (Disjunctive extension)
- If $K \circ \phi_1 = K \circ \phi_2$ and $K \circ \psi_1 = K \circ \psi_2$ then $K \circ (\phi_1 \vee \psi_1) = K \circ (\phi_2 \vee \psi_2)$ (Disjunctive equivalence)

Proposition 4 *Every selective revision operator defined from Kuratowski transformation functions satisfies* Disjunctive extension *and* Disjunctive equivalence*.*

3. Except the second postulate, *Identity*, whose proof is immediate.

Proof For *Disjunctive extension* suppose $K \circ (\phi \vee \psi) \vdash \neg\phi$, i.e., $K * f_K(\phi \vee \psi) \vdash \neg\phi$. Since $\phi \vdash f_K(\phi \vee \psi) \vdash f_K(\phi \vee \psi \vee \theta)$ by *implication* and *monotonicity* we can use the fact that $*$ is an AGM revision operator to deduce from this $K * f_K(\phi \vee \psi \vee \theta) \vdash \neg\phi$, i.e., $K \circ (\phi \vee \psi \vee \theta) \vdash \neg\phi$ as required.

For *Disjunctive equivalence* suppose $K \circ \phi_1 = K \circ \phi_2$ and $K \circ \psi_1 = K \circ \psi_2$. Then $K * f_K(\phi_1) = K * f_K(\phi_2)$ and $K * f_K(\psi_1) = K * f_K(\psi_2)$. Since $*$ is an AGM revision operator, this implies $K * (f_K(\phi_1) \vee f_K(\psi_1)) = K * (f_K(\phi_2) \vee f_K(\psi_2))$. Using this with the fact f_K satisfies *disjunctive distribution* gives $K * f_K(\phi_1 \vee \psi_1) = K * f_K(\phi_2 \vee \psi_2)$, i.e., $K \circ (\phi_1 \vee \psi_1) = K \circ (\phi_2 \vee \psi_2)$ as required. $\square$

Thus far, we have been concerned with Kuratowski transformation functions. The postulates we have introduced apply in the context of trust-sensitive revision due to the following result.

Proposition 5 *Let f_Π be a transformation function defined via a state partition Π. Then f_Π is a Kuratowski transformation function. Thus every trust-sensitive revision operator satisfies all the postulates listed in Propositions 3 and 4.*

Proof If f_Π is a transformation function defined via a state partition Π then we have $|f_\Pi(\phi)| = \Pi[\phi] = \bigcup\{\Pi(s) \mid s \models \phi\}$. The proof that f_Π is a Kuratowski transformation function is then straightforward. For example, if we let $\sim_\Pi$ denote the equivalence relation defined from Π, i.e., $s \sim_\Pi t$ iff $s \in \Pi(t)$, then *implication* and *idempotence* follow from the reflexivity and transitivity of $\sim_\Pi$ respectively, *falsity preservation* and *extensionality* hold immediately, while *disjunctive distribution* follows from the fact $\Pi[\phi \vee \psi] = \Pi[\phi] \cup \Pi[\psi]$. $\square$

What, then, are the properties of trust-sensitive revision that do not necessarily hold for all selective revision operators defined by Kuratowski transformation functions? The following result introduces a new postulate that holds for trust-sensitive revision.

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)*

Proof For each K, let Π_K be the state partition associated to K. Let $S_1, \ldots, S_m$ be the cells of the partition Π_K, and then for each $i = 1, \ldots, m$ let λ_i be any formula whose models are precisely those in S_i. Note that, for any $i \neq j$, we have $f_K(\lambda_i) \equiv \lambda_i$ and $\lambda_i \wedge \lambda_j \vdash \bot$.
<u>*(i)* $(K \circ \lambda_i) \cup (K \circ \lambda_j) \vdash \bot$ for $i \neq j$</u> For any i we have $K \circ \lambda_i = K * f_K(\lambda_i) = K * \lambda_i$. Hence $K \circ \lambda_i \vdash \lambda_i$ and so, for $i \neq j$, $(K \circ \lambda_i) \cup (K \circ \lambda_j) \vdash \lambda_i \wedge \lambda_j \vdash \bot$.
<u>*(ii)* for all ϕ there exists a set $X \subseteq \{1, \ldots, m\}$ such that $K \circ \phi = \bigcap_{i \in X} K \circ \lambda_i$</u> We have $K \circ \phi = K * f_K(\phi)$. Let $Y = \{i \mid \phi \wedge \lambda_i \text{ is consistent}\}$. Then $f_K(\phi) \equiv \bigvee_{i \in Y} \lambda_i$ and so $K * f_K(\phi) = K * \bigvee_{i \in Y} \lambda_i$. By *Disjunctive factoring*, which holds for any AGM revision

operator $*$, we know there exists $X \subseteq Y$ such that $K * \bigvee_{i \in Y} \lambda_i = \bigcap_{i \in X} K * \lambda_i$, from which the result follows. $\square$

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. The following result shows that this postulate does not hold in general for selective revision operators defined via Kuratowski transformation functions.

Proposition 7 *There exists a Kuratowski transformation function f such that, for any AGM revision operator $*$, the operator $\circ$ defined via f and $*$ does not satisfy* Disjoint outcome basis*.*

Proof Assume $\mathbf{F} = \{p\}$ and define f as follows: $f(p) = f(\top) \equiv \top$, $f(\neg p) \equiv \neg p$, $f(\bot) \equiv \bot$. (Every formula in this language is equivalent to precisely one of $p, \neg p, \top, \bot$, so these four values completely specify f by appeal to *extensionality*). One can check that f forms a Kuratowski transformation function. Let $K_\top = Cn(\top)$ and let $*$ be any AGM revision operator. Then for all ϕ we have $K_\top * \phi = Cn(\phi)$, so if $\circ$ is defined via $f, *$ we have $K_\top \circ \phi = Cn(f(\phi))$. Hence the set of possible outcomes for $K_\top \circ \phi$ is $\{Cn(\top), Cn(\neg p), Cn(\bot)\}$. Assume for contradiction that *Disjoint outcome basis* holds. It says there exists some subset X of this set of possible outcomes such that *(i)* the elements of X are mutually inconsistent, and *(ii)* each individual possible outcome for $K_\top \circ \phi$ can be expressed as an intersection over some of the elements of X. Since $Cn(\top)$ cannot be expressed as an intersection over any subset of the other two possible outcomes, we must have $Cn(\top) \in X$. Similarly $Cn(\neg p)$ cannot be expressed in terms of the other two outcomes, so $Cn(\neg p) \in X$. But then X contains two mutually consistent elements, namely $Cn(\top)$ and $Cn(\neg p)$, contradicting *(i)*. Hence *Disjoint outcome basis* cannot hold. $\square$

Are there any more postulates, beyond those of Propositions 3, 4 and 6, that we obtain by requiring f_K to be derived from a state partition? To answer this, it is helpful to first ask another question: what properties on f_K, beyond the five stated just before Definition 8, will f_K derived from a state partition satisfy? In fact, within the class of Kuratowski transformation functions, the following property characterises the class of f that are derived from state partitions.

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

For those familiar with topology, this can be explained in terms of the basic concepts of *open* and *closed* sets. In topological terms, *Pawlak* essentially says that the complement of a closed set of states is itself closed, i.e., every open set is closed. The name is chosen since adding this property to the Kuratowski properties leads us to the class of *upper approximation* functions, as introduced by Pawlak (1982), building on the work of Yao (1998).

Proposition 8 (Yao, 1998) *The following are equivalent:*
(i). *$f = f_\Pi$ for some state partition Π.*
(ii). *f is a Kurarotowski transformation function that satisfies* Pawlak*.*

The *Pawlak* property can be given a natural interpretation in terms of trust in an agent B. Suppose, when receiving information ϕ from B, agent A wants to use a Kuratowski transformation function f_B to first filter out that part of ϕ over which A does not perceive B to have expertise. If $f_B(\phi) \equiv \phi$ then A accepts formula ϕ from B at face value and then incorporates it into its belief set. In this case we might say *A trusts B on formula ϕ*. In the presence of the Kuratowski properties, one can show that *Pawlak* is equivalent to the following:

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

Thus *Pawlak* corresponds to the property that A trusts B on ϕ iff A trusts B on $\neg\phi$. In other words, the relation of trust in B over a given formula is *invariant* under taking negations. Such an axiom of *symmetric trust* has already been suggested by Liau (2003, Axiom C3).

Another interesting property of any f derived from a state partition (which crops up in the proof of soundness of the *Deficit makeup* postulate below) is the following:

- $f(\phi) \wedge f(\psi) \vdash f(\phi \wedge f(\psi))$ (conjunctive weakening)

Soundness of this property for f_Π requires the symmetry and transitivity of $\sim_\Pi$, the equivalence relation that defines the partition Π. In fact, in the presence of the five Kuratowski properties, it can be used to *derive Pawlak*, and thus could be used as an alternative to *Pawlak* as the property that characterises the f derived from state partitions within the class of Kuratowski transformation functions.

Armed with this understanding of the behaviour of the transformation functions derived from state partitions, we can now prove the soundness of another postulate for the family of trust-based revision operators.

- If $K \circ \phi \subseteq K \circ \psi$ there exists θ such that $\theta \vdash \phi$ and $K \circ \theta = K \circ \psi$. (Deficit makeup)

Note that *Deficit makeup* naturally holds for AGM revision. In that case we can take $\theta = \phi \wedge \psi$.

Proposition 9 *Every trust-sensitive belief revision operator satisfies* Deficit makeup.

Proof In what follows, we assume $f_K = f_\Pi$ for some state partition Π.

Suppose $K \circ \phi \subseteq K \circ \psi$, i.e., $|K \circ \psi| \subseteq |K \circ \phi|$, which means $\min_{\prec_K}(|f_K(\psi)|) \subseteq \min_{\prec_K}(|f_K(\phi)|)$. Let $\theta = \phi \wedge f_K(\psi)$. Clearly $\theta \vdash \phi$, so it remains to show $K \circ \theta = K \circ \psi$, equivalently $|K \circ \theta| = |K \circ \psi|$. We prove each inclusion of the latter in turn.

To show $|K \circ \theta| \subseteq |K \circ \psi|$, let $x \in |K \circ \theta|$, i.e., $x \in \min_{\prec_K}(|f_K(\theta)|)$. We must show $x \in \min_{\prec_K}(|f_K(\psi)|)$. From $x \in |f_K(\theta)|$ we know $x \in |f_K(\phi \wedge f_K(\psi))|$. Since $f_K(\phi \wedge f_K(\psi)) \vdash f_K(\psi)$ using the *monotonicity* and *idempotence* properties of f_K we thus know $x \in |f_K(\psi)|$. It remains to show x is $\prec_K$-minimal in $|f_K(\psi)|$. Assume not, for contradiction. Then $u \prec_K x$ for some $u \in \min_{\prec_K}(|f_K(\psi)|)$. By the initial assumption $|K \circ \psi| \subseteq |K \circ \phi|$ this means $u \prec_K x$ for some $u \in \min_{\prec_K}(|f_K(\phi)|)$. From $u \in |f_K(\phi)| \cap |f_K(\psi)|$ we know $u \in |f_K(\phi \wedge f_K(\psi))|$, i.e., $u \in |f_K(\theta)|$, from the *conjunctive weakening* property of f_K. Hence we have $u \prec_K x$ and $u \in |f_K(\theta)|$ - contradicting $x \in \min_{\prec_K}(|f_K(\theta)|)$. Hence x is $\prec_K$-minimal in $|f_K(\psi)|$ as required.

To show $|K \circ \psi| \subseteq |K \circ \theta|$, let $x \in |K \circ \psi|$, i.e., $x \in \min_{\prec_K}(|f_K(\psi)|)$. By the assumption $|K \circ \psi| \subseteq |K \circ \phi|$ we also have $x \in \min_{\prec_K}(|f_K(\phi)|)$ so $x \in |f_K(\phi)| \cap |f_K(\psi)| \subseteq |f_K(\phi \wedge f_K(\psi))| = |f_K(\theta)|$ by *conjunctive weakening*. It remains to show x is $\prec_K$-minimal in $|f_K(\theta)|$. Suppose, for contradiction, $y \prec_K x$ and $y \in |f_K(\theta)|$. But $f_K(\theta) = f_K(\phi \wedge f_K(\psi)) \vdash f_K(\psi)$ by *monotonicity* and *idempotence*. Hence $y \in |f_K(\psi)|$ - contradicting $x \in \min_{\prec_K}(|f_K(\psi)|)$. Hence x is $\prec_K$-minimal in $|f_K(\theta)|$, as required. $\square$

5. Representation Result

The previous section provided a list of postulates that were sound for any selective revision operator defined from an AGM revision operator and a state partition. The purpose of this section is to characterise this class of selective revision operators, and thereby precisely pin down the properties of trust-sensitive revision $*^B_A$. Our representation result is as follows:

Theorem 1 *Let $\circ$ be a function that, for any belief set K and formula ϕ, returns a belief set $K \circ \phi$. The following are equivalent:*
(i). $\circ$ *satisfies* Closure*,* Proxy success*,* Consistency*,* Extensionality*,* Disjunctive inclusion*,* Disjunctive factoring*,* Disjunctive extension*,* Disjunctive equivalence *and* Deficit makeup*.*
(ii). *There exists an AGM revision operator $*$ (i.e., a faithful assignment) and a state partition Π_K for each belief set K such that, for all K, ϕ, $K \circ \phi = K * f_{\Pi_K}(\phi)$*

The soundness part *(ii) $\Rightarrow$ (i)* has been proved in propositions 5 and 9. To show the completeness part *(i) $\Rightarrow$ (ii)* we need to define, from a given revision operator $\circ$, a faithful assignment $K \mapsto \preceq_K$ and a state partition assignment $K \mapsto \Pi_K$. So in the remainder of this section we assume $\circ$ satisfying the postulates in *(i)* is given. Then, for each K, we define $\preceq_K$ by setting, for each $x, y \in 2^{\mathbf{F}}$,[4]

$$x \preceq_K y \text{ iff } K \circ y \vdash \neg y \text{ or } [K \circ x \nvdash \neg x \text{ and } K \circ (x \vee y) \subseteq K \circ x].$$

The partition Π_K is defined via its associated equivalence relation $\sim_K$ as follows:

$$x \sim_K y \text{ iff } K \circ x = K \circ y.$$

We need to show $\preceq_K$ is a total preorder. (Note in what follows we sometimes omit explicit mention of the more obvious applications of the postulate *Extensionality*.)

Lemma 1 *$\preceq_K$ is a total preorder over $2^{\mathbf{F}}$.*

Proof We must show completeness and transitivity.
Completeness: $x \preceq_K y$ or $y \preceq_K x$. If $K \circ y \vdash \neg y$ or $K \circ x \vdash \neg x$ then we are done, so assume both $K \circ y \nvdash \neg y$ and $K \circ x \nvdash \neg x$. Then we need to show either $K \circ (x \vee y) \subseteq K \circ x$ or $K \circ (x \vee y) \subseteq K \circ y$. But this follows from *Disjunctive factoring*.
Transitivity: $[x \preceq_K y$ and $y \preceq_K z] \Rightarrow x \preceq_K z$. If $K \circ z \vdash \neg z$ then we are done, so assume $K \circ z \nvdash \neg z$. Then from $y \preceq_K z$ we know $K \circ y \nvdash \neg y$ and $K \circ (y \vee z) \subseteq K \circ y$. From $x \preceq_K y$

4. Note that, whenever a state x appears within the scope of a revision operator or of a logical connective, we are using it as shorthand for *prop*(x) (see Definition 5).

and $K \circ y \nvdash \neg y$ we get $K \circ x \nvdash \neg x$ and $K \circ (x \vee y) \subseteq K \circ x$. We must show $K \circ (x \vee z) \subseteq K \circ x$, equivalently (by *Disjunctive inclusion*) $K \circ (x \vee z) \nvdash \neg x$. By *Disjunctive extension* it suffices to prove $K \circ (x \vee y \vee z) \nvdash \neg x$. From $K \circ y \nvdash \neg y$ and $K \circ (y \vee z) \subseteq K \circ y$ we know $K \circ (y \vee z) \nvdash \neg y$ and similarly $K \circ (x \vee y) \nvdash \neg x$. Hence, by *Closure*, $\neg(x \vee y) \notin K \circ (y \vee z) \cup K \circ (x \vee y)$. Using *Disjunctive factoring* this gives $\neg(x \vee y) \notin K \circ (x \vee y \vee z)$, so $K \circ (x \vee y \vee z) \subseteq K \circ (x \vee y)$ by *Disjunctive inclusion*. From this and $\neg x \notin K \circ (x \vee y)$ we get $\neg x \notin K \circ (x \vee y \vee z)$ as required. □

Next we need to show $|K \circ \phi| = \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$. Before that we need another lemma.

Lemma 2 *Let $x, y \in 2^{\mathbf{F}}$. If $K \circ y \nvdash \neg x$ then $K \circ x = K \circ y$.*

Proof Assume $K \circ y \nvdash \neg x$. We first show $K \circ y \subseteq K \circ x$. By *Proxy success* there exists θ such that $K \circ y = K \circ (y \vee \theta) \vdash y \vee \theta$. We know $x \in |\theta|$ because otherwise $K \circ (y \vee \theta) \vdash y \vee \theta \vdash \neg x$, contradicting $K \circ y \nvdash \neg x$. Hence, by *Extensionality*, $K \circ y = K \circ (y \vee \theta) = K \circ (x \vee y \vee \theta')$ for some θ'. Hence, since $K \circ y \nvdash \neg x$, $K \circ (x \vee y \vee \theta') \nvdash \neg x$ so $K \circ (x \vee y \vee \theta') \subseteq K \circ x$ by *Disjunctive inclusion*, i.e., $K \circ y \subseteq K \circ x$ as required.

Now, to prove $K \circ x = K \circ y$, we know from the above-proved $K \circ y \subseteq K \circ x$ together with *Deficit makeup* that there exists some θ such that $\theta \vdash y$ and $K \circ \theta = K \circ x$. If $\theta \vdash y$ then either $\theta \equiv \perp$ or $\theta \equiv y$, hence we must have either $K \circ \perp = K \circ x$ or $K \circ y = K \circ x$. But, by *Consistency*, $K \circ \perp$ is inconsistent, while $K \circ x$ is consistent. Hence it cannot be the case that $K \circ \perp = K \circ x$, leaving us with $K \circ y = K \circ x$ as required. □

We now take each inclusion direction of $|K \circ \phi| = \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$ in turn.

Lemma 3 *For all K, ϕ, we have $|K \circ \phi| \subseteq \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$.*

Proof Let $x \in |K \circ \phi|$. We need to show *(i)* $x \sim_K y$ for some $y \in |\phi|$, and *(ii)* $x \preceq_K z$ for all z such that $z \sim_K y$ for some $y \in |\phi|$.

To show *(i)* we must show $K \circ x = K \circ y$ for some $y \in |\phi|$. If $x \in |\phi|$ then we are done, so assume $x \notin |\phi|$. By *Disjunctive factoring* $x \in |K \circ y|$ for some $y \in |\phi|$. Then $K \circ y \nvdash \neg x$. Then $K \circ x = K \circ y$ by Lemma 2.

To show *(ii)* assume $x \in |K \circ \phi|$. Let $K \circ z = K \circ y$, where $y \in |\phi|$. We must show $x \preceq_K z$, i.e., either $K \circ z \vdash \neg z$ or $[K \circ x \nvdash \neg x$ and $K \circ (x \vee z) \subseteq K \circ x]$. If $K \circ z \vdash \neg z$ we are done, so assume $K \circ z \nvdash \neg z$, i.e., $K \circ y \nvdash \neg z$. Now we must show 2 things: *(a)* $K \circ x \nvdash \neg x$, and *(b)* $K \circ (x \vee z) \subseteq K \circ x$. To show *(a)*, from *Proxy success* $K \circ \phi = K \circ \psi \vdash \psi$ for some ψ. Hence from $x \in |K \circ \phi|$ we know $K \circ \psi \nvdash \neg x$ and $x \in |\psi|$. From this and *Extensionality* we get $K \circ (x \vee \psi) \nvdash \neg x$. Then from *Disjunctive extension* and *Extensionality* $K \circ x \nvdash \neg x$ as required. To show *(b)*, from $K \circ z = K \circ y$ we get $K \circ (x \vee z) = K \circ (x \vee y)$ from *Disjunctive equivalence*. If we can show $K \circ (x \vee y) \nvdash \neg x$ then $K \circ (x \vee z) \nvdash \neg x$ so $K \circ (x \vee z) \subseteq K \circ x$ as required, by *Disjunctive inclusion*. So suppose for contradiction $K \circ (x \vee y) \vdash \neg x$. Then $K \circ (x \vee \bigvee_{y' \in |\phi|} y') \vdash \neg x$ by *Disjunctive extension*. Hence $K \circ (x \vee \phi) \vdash \neg x$ by *Extensionality*. By *Disjunctive factoring* $K \circ (x \vee \phi)$ equals either $K \circ x$ or $K \circ x \cap K \circ \phi$ or $K \circ \phi$. In the first two cases $K \circ x \vdash \neg x$ – contradicting part *(a)* proved above. In the third case $K\circ\phi \vdash \neg x$ – contradicting the assumption. Hence $K\circ(x\vee y) \nvdash \neg x$ as required. $\square$

Before proving the converse direction we need one more lemma.

Lemma 4 *If* $x \in \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$ *then* $K \circ x \nvdash \neg x$.

Proof Assume for contradiction $K\circ x \vdash \neg x$. Then, for all $y \in \Pi_K[\phi]$, from this and $x \preceq_K y$ we must have $K\circ y \vdash \neg y$. Hence $K\circ \bigvee_{z\in\Pi_K[\phi]} z \vdash \neg y$ for all $y \in \Pi_K[\phi]$ by *Disjunctive extension.* Hence $K\circ\bigvee_{z\in\Pi_K[\phi]} z \vdash \neg \bigvee_{y\in\Pi_K[\phi]} y$. But we already know $K\circ\bigvee_{z\in\Pi_K[\phi]} z \vdash \bigvee_{y\in\Pi_K[\phi]} y$ from Lemma 3 above, thus giving the required contradiction. $\square$

Given this result, we can now prove the converse to Lemma 3.

Lemma 5 *For all* K, ϕ, *we have* $\min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\}) \subseteq |K \circ \phi|$.

Proof Let $x \in \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$, i.e., $K \circ x = K \circ z$ for some $z \in |\phi|$ and $x \preceq_K y$ for all $y \in \Pi_K[\phi]$. To prove the lemma, we must show $K \circ \phi \nvdash \neg x$, equivalently (by *Extensionality*) $K \circ (z_1 \vee \ldots \vee z_m) \nvdash \neg x$ where $|\phi| = \{z_1, \ldots, z_m\}$. Assume $K \circ x = K \circ z_1$. Then, by *Disjunctive equivalence* it suffices to show $K \circ (x \vee z_2 \vee \ldots \vee z_m) \nvdash \neg x$. This will be proved (by *Disjunctive factoring*) if we can show $K \circ (x \vee z_i) \nvdash \neg x$ for all $i = 2, \ldots, m$. So let $i \in \{2, \ldots, m\}$ and assume for contradiction $K \circ (x \vee z_i) \vdash \neg x$. Choose any $y \in 2^{\mathbf{F}}$ such that $K \circ z_i \nvdash \neg y$. (Such y exists by *Consistency*.) Then $K \circ z_i = K \circ y$ by Lemma 2. Hence $y \in \Pi_K[\phi]$. We now claim $y \prec_K x$, which gives us the required contradiction since $x \in \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$. To show $y \prec_K x$ we need to show *(a)* $K \circ (x \vee y) \nvdash \neg y$ and *(b)* $K \circ (x \vee y) \vdash \neg x$.

Part *(b)* follows from $K \circ (x \vee z_i) \vdash \neg x$, $K \circ z_i = K \circ y$ and *Disjunctive equivalence*. For part *(a)* we know $K \circ x \nvdash \neg x$ from $x \in \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$ and Lemma 4. Hence from this and $K \circ (x \vee z_i) \vdash \neg x$ we get $K \circ (x \vee z_i) = K \circ z_i$ by *Disjunctive factoring.* Hence, since $K \circ z_i \nvdash \neg y$ we get $K \circ (x \vee z_i) \nvdash \neg y$. But $K \circ z_i = K \circ y$, so we conclude $K \circ (x \vee y) \nvdash \neg y$ from *Disjunctive equivalence.* $\square$

Putting lemmas 3 and 5 together we then have that, for all K, ϕ, $|K \circ \phi| = \min_{\prec_K}(\{s \mid s \in \Pi_K[\phi]\})$. This concludes the proof that *(i)* $\Rightarrow$ *(ii)* holds in Theorem 1, and thus we have our representation result.

6. The Success Postulate

As we have seen, *Success* does not hold in general for trust-sensitive revision $*^B_A$. 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)

As an extreme counterexample to $*^B_A$ failing *Weak Success*, just take $T_A(K, B)$ to be the trivial partition so that $K *^B_A \phi$ 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 $f_K(\phi) \equiv \phi$ whenever $K \not\vdash \neg\phi$. 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 $K \not\vdash \neg\phi$ and $K \circ \neg\phi \vdash \neg\phi$ then $K \circ \phi \vdash \phi$ (Very Weak Success)

This rule roughly says that, if we accept B's information when it tells us ϕ is false, then we should accept B's information 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 $K \not\vdash \neg\phi$ and $K \not\vdash \phi$ then $[K \circ \phi \vdash \phi$ iff $K \circ \neg\phi \vdash \neg\phi]$ (Negation invariance)

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

Proposition 10 *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: $K \not\vdash \neg p$, $K \circ \neg p = Cn(\neg p \wedge q)$, so $K \circ \neg p \vdash \neg p$. However, $K \circ p = Cn(q)$, so $K \circ p \not\vdash p$.

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 acccept a report that p is false, but we will not accept a report that it is true.

We can frame Example 4 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. In this case p is true just in case the first battery works and q is true just in case the second battery works. Now suppose you contribute the battery corresponding to q and your adversary contributes the battery corresponding to p. Your belief state $K = Cn(q)$ captures the fact that you believe your battery works while you are unsure whether the other battery works or not. If your adversary tests the lamp in private and tells you that their battery works, we argue that you should not accept this conclusion. From your perspective, they may have seen the dim light and jumped to the conclusion that their battery is working. So, your adversary's report is p, but we do not want $K \circ p \vdash p$. On the other hand, had your adversary reported $\neg p$, we would want $K \circ \neg p \vdash \neg p$ because you would be entirely willing to believe that the battery they have contibuted does not work. So while you would accept a report that their battery is dead, you will not accept a report that it works. This is exactly what is captured in Example 4.

Inuitively, the asymmetric treatment of p and $\neg p$ in the example is due to the fact that each agent knows that the other may interpret observations differently. If the adversary initially believes p, and then observes

$$(p \wedge \neg q) \vee (\neg p \wedge q)$$

then they will continue to believe p, and they may announce it as true. However, if we initially believe $\neg p$, then this same observation will not convince us to believe p. So it does not make sense to trust p following a report of p from someone that already believed it to be true. On the other hand, if the adversary initially believes p, the only observation that could make them believe $\neg p$ would be that there is no light at all:

$$\neg p \wedge \neg q.$$

This report will convince any agent to believe $\neg p$. Hence, the asymmetry is caused by the fact that the observation is only seen by one agent who then reports their own beliefs. It does not make sense to blindly trust such a report, since it is influenced not only by a new observation, but also by the initial beliefs of the adversary.

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

Proposition 11 *Every trust-sensitive revision operator $*_A^B$ satisfies the following postulate:*

- *If $K \not\vdash \neg\phi$ 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)*

Proof Assume $K \not\vdash \neg\phi$ and $K \circ \neg\phi \vdash \neg\phi$. Then $K \neq K \circ \neg\phi$ and so $f(\neg\phi) \not\equiv \top$ (since otherwise $K = K * \top = K \circ \neg\phi$). Let $\psi = \neg f(\neg\phi)$. Since $f(\neg\phi) \not\equiv \top$ we know ψ is consistent. It remains to show $\psi \vdash \phi$ and $K \circ \psi \vdash \phi$. The former follows from the fact $\neg\phi \vdash f(\neg\phi)$ (by *implication*). To show the latter we have $K \circ \psi = K * f(\psi) \vdash f(\psi)$. But $f(\psi) = f(\neg f(\neg\phi)) \vdash \neg f(\neg\phi) = \psi$ (by *Pawlak*). Hence $K \circ \psi \vdash \psi \vdash \phi$ as required. $\square$

Feeble Success relaxes *Very Weak Success* by saying that if we accept B's report 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 ($K \circ p \not\vdash p$), 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 essential use of the *Pawlak* property of f. Indeed *Feeble Success* is a property not shared by every selective revision operator defined via a Kuratowski function.

Proposition 12 *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 $K \not\vdash \neg p$ and $K \circ \neg p = Cn(\neg p) \vdash \neg p$, thus the antecedent holds.

7. Manipulability

Next we consider a concept that has been extensively studied in areas such as voting theory, preference aggregation and belief merging (Chopra, Ghose, & Meyer, 2006; Everaere, Konieczny, & Marquis, 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 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 $K \circ \phi \not\vdash \phi$ and ψ is consistent then $K \circ \psi \not\vdash \phi$ (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, Fermé, Cantwell, & Falappa, 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 13 *If a revision operator $\circ$ satisfies both* Feeble Success *and* Non-Manipulability *then it satisfies* Very Weak Success*.*

Proof Suppose $\circ$ satisfies *Feeble Success* and *Non-Manipulability* and assume $K \nvdash \neg\phi$ and $K \circ \neg\phi \vdash \neg\phi$. We must show $K \circ \phi \vdash \phi$. By the assumptions and *Feeble Success* there exists a consistent formula ψ such that $\psi \vdash \phi$ and $K \circ \psi \vdash \phi$. Using the facts that ψ is consistent and $K \circ \psi \vdash \phi$ gives us $K \circ \phi \vdash \phi$ by *Non-Manipulability* as required. $\square$

Since we have already seen that trust-sensitive revision satisfies *Feeble Success* but *not*, in general *Very Weak Success* (Propositions 10 and 11), we can immediately state:

Proposition 14 *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.

There is a weaker version of *Non-manipulability* that does hold for any $\circ$ based on a Kuratowski transformation function:

- If $K \circ \phi \not\vdash \phi$ then $K \circ (\phi \vee \psi) \not\vdash \phi$ (Weak non-manipulability)

Proposition 15 *Every selective revision operator based on a Kuratowski transformation function satisfies* Weak non-manipulability*.*

Proof Suppose $K \circ \phi \not\vdash \phi$, i.e., $\min_{\prec_K}(|f(\phi)|) \not\subseteq |\phi|$. Then there exists $x \in |\neg\phi| \cap \min_{\prec_K}(|f(\phi)|)$. By *monotonicity*, from $x \in \min_{\prec_K}(|f(\phi)|)$ we know $x \in |f(\phi \vee \psi)|$. Suppose for contradiction $K \circ (\phi \vee \psi) \vdash \phi$. Then $x \in |\neg\phi|$ implies $x \notin \min_{\prec_K}(|f(\phi \vee \psi)|)$ so there exists $y \in \min_{\prec_K}(|f(\phi \vee \psi)|)$ such that $y \prec_K x$. From $K \circ (\phi \vee \psi) \vdash \phi$ we know $y \in |\phi|$ and so (since $\phi \vdash f(\phi)$ by *implication*) $y \in |f(\phi)|$. So we have shown $x \in \min_{\prec_K}(|f(\phi)|)$, $y \prec_K x$ and $y \in |f(\phi)|$ - contradiction. Hence $K \circ (\phi \vee \psi) \not\vdash \phi$ as required. $\square$

We note that *Weak non-manipulability* does not hold for general selective revision operators. For assume $\mathbf{F} = \{p, q\}$, with initial preorder $\prec_K$ given by

$$\emptyset \prec_K \{p\} \prec_K \{q\} \sim_K \{p, q\}.$$

Then assume $f(p) = \top$ and $f(p \vee q) = p \vee q$ (note f doesn't satisfy *monotonicity*, so is not a Kuratowski transformation function). We get $K \circ p = Cn(\neg p \wedge \neg q)$ and $K \circ (p \vee q) = Cn(p \wedge \neg q)$. Thus $K \circ p \not\vdash p$ but $K \circ (p \vee q) \vdash p$.

8. Discussion

There have been some related works from the literature on trust and on belief revision.

8.1 Related Models of Trust

Many different models of trust have been explored in the the literature, including work on reputation systems (Huynh, Jennings, & Shadbolt, 2006), task allocation (Ramchurn, Mezzetti, Giovannucci, Rodriguez-Aguilar, Dash, & Jennings, 2009), and deception (Salehi-Abari & White, 2009). One notable approach with an emphasis on knowledge representation has been developed by Wang and Singh (2007), who show how trust can be built based on evidence; this could be used as a precursor step to build a trust relation in our framework. Different levels of trust have been addressed by Krukow and Nielsen (2007), by using a lattice structure to represent various degrees 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.

The relationship between trust and belief change has been explored in Dynamic Epistemic Logic. One notable framework is the logic 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 (Lorini, Jiang, & Perrussel, 2014). 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*. However, it is possible to define trust-sensitive revision based on state partitions in the setting of Dynamic Epistemic Logic, and the resulting logic differs from DL-BT in a non-trivial manner. We leave the exploration of this connection for future work.

Alternative modal approaches to trust have appeared in the literature. For instance, Liau (2003) introduces a modal operator T_{ij} where $T_{ij}\phi$ is true just in case i trusts j on the truth of the proposition ϕ. The semantics is given by a relation $\mathcal{T}_{ij}$ that essentially maps each state s to the set of formulas over which j is trusted in s. Hence, in this approach, the set of formulas over which j is trusted is *explicit*. Our approach is different in that the set of formulas over which an agent is trusted is *implicit*; it is determined by the partition on states. The rationale for this distinction is the fact that we are interested in the influence of trust on belief change in the AGM tradition. Just as the total pre-orders underlying a revision operator are implicit to an agent in this setting, so too are the state partitions defining trust.

Another epistemic logic approach to trust is introduced by Rodenhäuser (2014). In this 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 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 by Hannson et al. (2001).

A distinct modal approach to trust has been defined in the context of speech act theory (Herzig & Longin, 2000). This work actually does not explicitly mention trust, referring instead to the notion of *competency*. The end result is the same: the formalism indicates when one agent should incorporate the information provided by another. In this case, however, the notion of a *topic* is used as an intermediary. Every agent is an authority over certain topics, and every formula is associated with certain topics. This allows us to determine when an agent should be believed on a formula. However, the association between formulas and topics is not extensional. In other words, logically equivalent formulas need not be associated with the same topics. This is quite different from our approach, in which logically equivalent formulas must be treated identically since trust is defined at the level of states.

8.2 Language Splitting in Belief Revision

In our motivating Example 2 in §3.1 the domains of expertise of the doctor D and the jeweler J could be neatly characterized in terms of the propositional symbols whose truth-values they could be trusted to discern: *sick* in the case of D and *diam* in the case of J. When receiving information $sick \wedge \neg diam$ from D, agent A is then able to "separate out" that part of the input pertaining to the symbol over which D is not perceived to have expertise, before incorporating that part of the input that remains.

The idea of splitting information into separate compartments over distinct sublanguages has cropped up periodically in the belief revision literature, usually in connection with the idea of *relevance*. It was first suggested by Parikh (1999), who introduced the idea of a ϕ-*splitting*, for a given propositional formula ϕ. A ϕ-splitting is a partition of the propositional symbols (not directly a partition of the states, as in our setting) such that ϕ is logically equivalent to a conjunction of formulas, with the propositional symbols of each conjunct being restricted to one cell of the partition. Parikh proved that every formula ϕ has a unique *finest* ϕ-splitting. For example, assuming $\mathbf{F} = \{p, q\}$, the finest $(p \wedge q)$-splitting is $\{\{p\}, \{q\}\}$, while the finest $(p \vee q)$-splitting is $\{\{p, q\}\}$.

Parikh's motivation for using splittings is as a way to compartmentalize the *current* belief set K into information about separate domains, and thus to localize the effects of revision to that part of K that shares the same domain as the new information. However, one could imagine employing the same idea to model domains of expertise in our setting.[5] The idea would be to assign, to each source B, the set of symbols $S(B) \subseteq \mathbf{F}$ over which A perceives B to have expertise, and then to define a transformation $f_{S(B)}(\phi)$ of input formula ϕ as follows: *(i)* reformulate ϕ as a conjunction over the finest ϕ-splitting, *(ii)* strike out all those conjuncts that contain variables *not* in $S(B)$. Indeed, this description seems to match perfectly with what happens in Example 2.

We note, however, that this division of trust strictly along the lines of the propositional symbols fails to capture some intuitive scenarios that are handled by our approach. In particular, when receiving input $p \wedge q$, the transformed formula $f_{S(B)}(p \wedge q)$ can only be equivalent to one of p, q, $p \wedge q$ or $\top$. But one can easily imagine a scenario in which B could be trusted to discern whether *at least one* of p, q are true, without necessarily being able to distinguish which one or how many, in which case it would seem reasonable to weaken B's information $p \wedge q$ to $p \vee q$. In our setting this could easily be captured by using f_Π for an appropriate state partition Π. Despite their restrictive nature, the class of selective revision operators obtained from such transformation functions $f_{S(B)}$ is still worthy of study, which we leave to another occasion.

Another work in the belief change literature that, similarly to Parikh, is motivated by trying to localize the effects of a revision only to that part of the belief state that is relevant to the incoming formula, is introduced by Hannson and Wassermann (2002). Like Parikh, and unlike us, they are concerned with compartmentalizing the current beliefs rather than

5. The idea of modeling domain expertise along the dimensions of the propositional symbols is also briefly mentioned by Liau (2003).

the input, and they work directly at the level of formulas rather than states. Unlike Parikh, these compartments are typically overlapping (since a currently believed formula can be relevant to different input formulas). However a major difference with our work is that, unlike in AGM, they represent belief states by *belief bases*, i.e., sets of formulas that are not closed under logical consequence.

8.3 Future Work

There are many directions for future work. One direction would be to explore the properties of trust-sensitive revision in the setting of Dynamic Epistemic Logic; the interaction between trust and belief in this setting is more complex than the in propositional case. Another direction would be to explore 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 in the context of 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.

One additional direction for future work is related to the notion of *strength of trust*. This is particularly important in cases where mulitiple agents provide conflicting reports. In the remainder of this section, we outline a general notion of strength of trust that we will expand in future work.

In order to capture different levels of trust, we can introduce a measure of the distance between states. Each agent A will associate a distance function $d_{B,K}$ over states with each agent B and belief set K. If $d_{B,K}(s,t) = 0$, then B cannot be trusted to distinguish between the states s and t. If $d_{B,K}(s,t)$ 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.

We propose pseudo-ultrametrics for measuring distance due to the following basic property from the theory of metric spaces. Given a pseudo-ultrametric d over a set X, an element $x \in X$ and a fixed number n, define:

$$rad(x,n) = \{y \mid d(x,y) \leq n\}.$$

For any fixed n, the collection of sets $\{rad(x,n) \mid x \in X\}$ is a partition. Hence, a pseudo-ultrametic can be used to define a sequence of partitions.

If we use pseudo-ultrametrics to represent trust, the preceding property suggests an approach through which strength of trust can be used to resolve conflicting reports from multiple agents. In particular, if B and C offer conflicting reports, we can look at the sequence of trust partitions obtained from the distance functions $d_{B,K}$ and $d_{C,K}$ associated with B and C, respectively. There will be some least n such that the intersection of the n^{th} partitions is consistent; we can then define this as the joint partition to be used for revision. This approach guarantees that the agent that is trusted "more" on a particular distinction will be believed when there is a conflict, because as the partitions get coarser, we are essentially weakening the information to be used in the revision. The procedure of incrementally weakening the formulas for revision until consistency is reached is highly reminiscent of the *belief negotiation* approach to belief merging (Booth, 2006; Konieczny & Pino Pérez, 2002). We leave the full development of this extension for future work.

8.4 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 have discussed many directions for future work, including extensions to handle strength of trust, iterated belief change, and modal operators. We suggest that our work could also be used in practical situations where agents need to make decisions based on reported information, including applications in Security and Network Communication.

Acknowledgements

Thanks are due to Emil Weydert for pointing us to the links with topology. Thanks also to the reviewers for some useful comments that helped to improve the paper.

References

Alchourrón, C., Gärdenfors, P., & Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, *50*(2), 510–530.

Booth, R. (2006). Social contraction and belief negotiation. *Information Fusion*, *7*(1), 19–34.

Chopra, S., Ghose, A., & Meyer, T. (2006). Social choice theory, belief merging, and strategy-proofness. *Information Fusion*, *7*(1), 61–79.

Everaere, P., Konieczny, S., & Marquis, P. (2007). The strategy-proofness landscape of merging. *Journal of Artificial Intelligence Research*, *28*, 49–105.

Fermé, E., & Hansson, S. O. (1999). Selective revision. *Studia Logica*, *63*(3), 331–342.

Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, *41*(4), 587–601.

Hansson, S. O. (1997). Semi-revision. *Journal of Applied Non-Classical Logics*, *7*(1-2), 151–175.

Hansson, S. O., Fermé, E., Cantwell, J., & Falappa, M. (2001). Credibility-limited revision. *Journal of Symbolic Logic*, *66*(04), 1581–1596.

Hansson, S. O., & Wassermann, R. (2002). Local change. *Studia Logica*, *70*(1), 49–76.

Herzig, A., & Longin, D. (2000). Belief dynamics in cooperative dialogues. *Journal of Semantics*, *17*(2), 91–115.

Hunter, A., & Booth, R. (2015). Trust-sensitive belief revision. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3062–3068.

Huynh, T. D., Jennings, N. R., & Shadbolt, N. R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, *13*(2), 119–154.

Katsuno, H., & Mendelzon, A. (1992). Propositional knowledge base revision and minimal change. *Artificial Intelligence*, *52*(2), 263–294.

Konieczny, S., & Pino Pérez, R. (2002). Merging information under constraints: A logical framework. *Journal of Logic and Computation*, *12*(5), 773–808.

Krukow, K., & Nielsen, M. (2007). Trust structures. *International Journal of Information Security*, *6*(2-3), 153–181.

Kuratowski, K. (1966). *Topology, Volume 1*. Academic Press.

Liau, C. (2003). Belief, information acquisition, and trust in multi-agent systems — A modal logic formulation. *Artificial Intelligence*, *149*, 31–60.

Lorini, E., Jiang, G., & Perrussel, L. (2014). Trust-based belief change. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI)*, pp. 549–554.

Parikh, R. (1999). Beliefs, belief revision, and splitting languages. *Logic, language and computation*, *2*(96), 266–268.

Pawlak, Z. (1982). Rough sets. *International Journal of Parallel Programming*, *11*(5), 341–356.

Ramchurn, S. D., Mezzetti, C., Giovannucci, A., Rodriguez-Aguilar, J. A., Dash, R. K., & Jennings, N. R. (2009). Trust-based mechanisms for robust and efficient task allocation in the presence of execution uncertainty. *Journal of Artificial Intelligence Research*, *35*, 119–159.

Rodenhäuser, B. (2014). *A Matter of Trust: Dynamic Attitudes in Epistemic Logic*. Ph.D. thesis, University of Amsterdam.

Salehi-Abari, A., & White, T. (2009). Towards con-resistant trust models for distributed agent systems. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 272–277.

Satterthwaite, M. (1975). 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.

Wang, Y., & Singh, M. P. (2007). Formal trust model for multiagent systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1551–1556.

Yao, Y. (1998). Constructive and algebraic methods of the theory of rough sets. *Information Sciences*, *109*(1), 21–47.