Trustworthiness by default

Johan W. Klüwer[1] and Arild Waaler[2]

[1] Dep. of Philosophy, University of Oslo johanw@filosofi.uio.no
[2] Finnmark College and Dep. of Informatics, University of Oslo, arild@ifi.uio.no.

> *But never put a person to death on the testimony of only one witness. There must always be at least two or three witnesses.*
> *Deuteronomy 17:6 (New Living Translation)*

Abstract. We present a framework for reasoning about trustworthiness, with application to conflict resolution and belief formation at various degrees of reliability. On the basis of an assignment of relative trustworthiness to sets of information sources, a lattice of degrees of trustworthiness is constructed; from this, a priority structure is derived and applied to the problem of forming the right opinion in the presence of possibly conflicting information. Consolidated with an unquestioned knowledge base, this provides an unambiguous account of what an agent should believe, conditionally on which information sources are trusted. Applications in multi-agent doxastic logic are sketched.

1 Introduction

To trust an information source, in the simplest, unconditional form, is to believe every piece of information that the source provides. While providing a paradigm, this notion of trust has limited application to realistic scenarios. In general, the trust we have in our information sources, which may vary in kind from teachers to newspapers to legal witnesses, is not unconditional: we believe what we are told by a trusted source only as long as we don't possess knowledge to the contrary. This simple observation motivates the approach to trust that we will be discussing in this paper. Conditional trust in an information source is a *default* attitude: To believe what you are told, unless you know better.

When looking for information, we often need to consider several sources. Sources may vary widely with regard to their reliability, and a cautious default approach then informs us to let the more trustworthy ones take priority over those that are less trustworthy. Furthermore, we often need to consider more than one source at a time. Notions of agreement or corroboration, as well as the consolidation of information drawn from different sources, are essential.

What we present here is a framework for reasoning about relative trustworthiness, with *sets* of information sources as the basic trusted units. The main part of the paper is structured as follows. Section 2 addresses properties of the trust relation itself, making only informal reference to notions of information. Building on a simple trustworthiness relation (2.1), rational trust attitudes are identified and ordered according to strength (2.2, 2.3), and ordered in a tree structure of "fallbacks" (2.4). Section 3 employs this structure to provide an account of trust in terms of default conditionals. Notions of information, as provided by individual sources as well as collections of sources, are defined in 3.1. The prioritized default logic $\text{Æ}_{\top}$ is briefly presented in section 3.2. The defaults approach is then made explicit in section 3.3, which presents a method for expressing trust attitudes as formulae of $\text{Æ}_{\top}$.

For the presentation of the core theory, we assume that the information provided by sources is expressed in propositional logic. However, the theory is equally applicable if one wants to use a more, or less, complex language. Looking forward, section 3.4 outlines how the analysis can be applied to multi-agent doxastic logic, to enable the representation of doxastic agents with varying degrees of trust that the beliefs of other agents are true.

The expression of trusting attitudes in terms of prioritized defaults provides an answer to the following non-trivial question: Given that we possess a body of antecedent knowledge, and are provided with information from a set of variously trusted sources, what is it reasonable to believe?

This work builds on two main sources. For the theory of trustworthiness, the most important is the work of John Cantwell [1, 2], in which the basic relation of trustworthiness is defined in a way that is close to the one given here. For the aspects that relate to default inference and belief, the prioritized belief logic Æ [9, 10, 12], which is closely related to that of [7], has been the primary source of reference.

We consider the following to be guiding principles for what follows.

Given a collection of sources, what all sources agree on is at least as trustworthy as what only some agree on. (1)

If some unit x is trusted, and y is at least as trustworthy as x, then rationality demands that y should be trusted too. (2)

Accept information from a trusted unit as true, unless it is inconsistent with what you have already accepted. (3)

2 A trustworthiness relation

2.1 The basic pre-order on information sources

Let $\mathfrak{S}$ be a (possibly empty) finite set of *sources*. The *trustworthiness relation* $\trianglelefteq$ is a relation between subsets of $\mathfrak{S}$; we will often refer to these as *source units*. A source unit is an entity that is capable of providing information, as follows: A singleton unit $\{a\}$ provides exactly what the single source a does. A non-singleton unit provides only what follows from the contribution of every member. Informally, think of a non-singleton source unit as making a "common statement", i.e., the strongest that its members all agree on.

Notation: Small Latin letters a, b, c denote sources, small variable letters x, y, z range over source units, capital Latin letters A, B, C denote particular sets of source units, and capital variable letters X, Y, Z range over arbitrary sets of source units. We will sometimes have to collect sets of source units, for which we shall use capital Greek letters Γ, Δ.

We assume that the trustworthiness relation is reflexive and transitive (a *pre-order*). Two source units x and y may be *trustworthiness-equivalent*, written $x \sim y$.

$$x \sim y \quad =_{\text{def}} \quad x \trianglelefteq y \text{ and } y \trianglelefteq x \tag{4}$$

We write $x \lhd y$ to express that y is strictly more trustworthy than x.

$$x \lhd y \quad =_{\text{def}} \quad x \trianglelefteq y \text{ and not } x \sim y \tag{5}$$

Source units that are unrelated by $\trianglelefteq$ will be called *independent*, denoted $x \wr y$. Intuitively, we interpret independence as a consequence of lack of knowledge; neither of $x \rhd y$, $x \lhd y$, and $x \sim y$ is known to obtain. If no two source units are independent, we say $\trianglelefteq$ is *connected*.

We assume that every source, however it is combined with other sources, makes a non-negative contribution of information. Together with (1), this implies that enlargement of a source unit with new members may never yield a unit that provides a stronger set of information. Hence, a unit will be at least as trustworthy as every unit that it contains as a subset. This motivates taking the following principle, which we will occasionally refer to as *monotonicity*, to be valid.

$$x \trianglelefteq x \cup y\,. \tag{6}$$

It follows that for each source unit x, the following hold.

$$x \trianglelefteq \mathfrak{S}\,, \tag{7}$$

$$\emptyset \trianglelefteq x\,. \tag{8}$$

To see why (7) is valid, note that $\mathfrak{S}$ only provides information which is common to, is agreed upon, by all the sources. At the other extreme, we stipulate that the empty set is a limit case that always provides inconsistent information, motivating (8).

In referring to particular source units in examples we will consistently simplify notation by omitting brackets: $a \lhd bc$ is, e.g., shorthand for $\{a\} \lhd \{b, c\}$. Likewise, the set $\{\{a\}, \{a, b\}\}$ will be denoted a, ab. Observe that the symbol a should, depending on the context, either be taken as a reference to the source a or to the singleton source set $\{a\}$ or to the singleton source set collection $\{\{a\}\}$.

2.2 The poset of trust-equivalent source units

To have an *attitude* of trust, given some $\mathfrak{S}$, is to trust a (possibly empty) set of source units. In the following, we will allow ourselves to talk about attitudes as being the sets of source units themselves, and to say that a source unit is “included” in an attitude of trust, meaning that that source unit is among those trusted. The empty set represents the attitude of placing trust in none of the sources.

Given a trust relation $\lhd$, we can distinguish those trust attitudes that respect the relation. The relevant principle is expressed in rule (2), that x may only be trusted if every $y \trianglerighteq x$ is trusted as well. We will in this section identify the *permissible* trust attitudes according to this principle.

We will use the following standard terminology. In a *poset* $(S, \leq)$ the $\leq$-relation is reflexive, transitive and anti-symmetric. The poset has a unique cover relation $\prec$, defined as $x \prec y$ iff $x < y$ and $x \leq z < y$ implies $z = x$. $C \subseteq S$ is an *antichain* if every two distinct elements in C are incomparable by $\leq$. Note in particular that $\emptyset$ is an antichain. Every subset of S has $\leq$-minimal elements, and the set of these elements is an antichain. $\uparrow C$ denotes an *up-set*, defined as $\{x \mid (\exists y \in C)(y \leq x)\}$. The set of antichains in a poset is isomorphic to the set of up-sets under set inclusion.

If an attitude of trust includes a source set x, but not an equivalently trustworthy source set y, then the attitude is not permissible. This motivates a focus on the equivalence classes of $\mathfrak{S}$ modulo $\sim$. Where $x \subseteq \mathfrak{S}$,

$$[x] \quad =_{\text{def}} \quad \{y : x \sim y\} \tag{9}$$

Let $\dot{\mathfrak{S}}$ be the set of all equivalence classes of $\mathfrak{S}$ modulo $\sim$. We will say a set of sources x is *vacuous* with regard to trustworthiness if $x \in [\emptyset]$. In the extreme case that every set of sources is a member of $[\emptyset]$, the trustworthiness relation itself is said to be vacuous.

Where X and Y are in $\dot{\mathfrak{S}}$, define a relation $\dot{\lhd}$ of relative strength between them as follows.

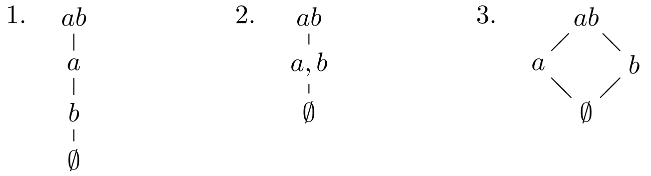
$$X \mathrel{\dot{\lhd}} Y \quad =_{\text{def}} \quad (\exists x \in X)(\exists y \in Y)(x \lhd y) \tag{10}$$

Let $X \mathrel{\dot{\trianglelefteq}} Y$ designate $X \mathrel{\dot{\lhd}} Y$ *or* $X = Y$ and let $X \mathrel{\dot{\wr}} Y$ designate independence.

Lemma 1. *$(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$ is a poset in which $[\emptyset]$ is the unique minimum and $[\mathfrak{S}]$ the unique maximum. $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$ is a linear order iff $(\wp\,\mathfrak{S}, \trianglelefteq)$ is connected.*

Proof. Monotonicity entails the unique minimum and maximum. The other properties follow easily from the construction of $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$.

Example 1. Assume that the set of sources $\mathfrak{S}$ contains just a and b, and that $a \lhd ab$, $b \lhd ab$, $\emptyset \lhd a$, and $\emptyset \lhd b$ (i.e., the source units a, b, and ab are non-vacuous, and ab is more trustworthy than both a and b). The following figure shows Hasse diagrams of the poset $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$, given 1. $a \lhd b$, 2. $a \sim b$, and 3. $a \wr b$.

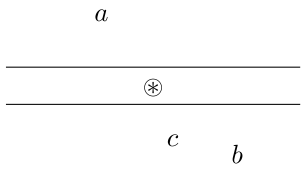


Relation 1. requires information provided by a to take precedence over information provided by b. Relation 2. emerges from taking a to precisely as reliable as b: it is only rational to accept a's contribution given that b's is accepted as well (in the event that a and b contradict each other, it is ruled out that either can be trusted separately). Relation 3. reflects a situation in which less is known about the relative trustworthiness of a and b than in 1. and 2, i.e., neither is known to be better or equivalent to the other. With this relation, trusting b but not a is not irrational; so the range of admissible attitudes is wider. In particular, where the information a provides is incompatible with what b provides, the relation doesn't rule out making a *choice* of trusting just one of the two.[1] Compared to 1. and 2., this relation offers more freedom, but less guidance.

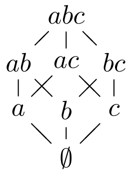
The following example, which is developed further in later sections, applies the theory to a reasonably realistic scenario.

Example 2 (Traffic accident). A traffic accident has occurred. We have been assigned the task of finding witnesses, assessing their relative trustworthiness, gathering their statements on what came to pass, weighing the evidence according to trustworthiness and finally presenting an account of the accident according to a reasonable standard (threshold) of trust.

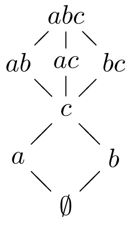
Assume, for this example, that the criterion according to which sources are deemed trustworthy or not is their viewpoint relative to the incident, and that we are provided with a drawing (right), illustrating the accident ⊛ and the positions of the witnesses. At the outset, we know that there are three witnesses, a, b, and c, but nothing about their respective trustworthiness. Making no prior assumptions, we start out with the weakest possible trust relation (0. below).



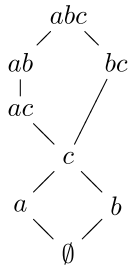
0.



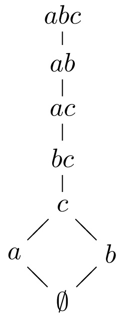
1.



2.



3.



By applying information provided by the drawing, we are able to considerably strengthen the trust relation. We will consider a sequence of three steps.

[1] When the case arises that a and b contradict each other, a choice will implicitly favour a revision of the trust relation to be like 1. or 2. If the subject opts to trust a over b, 1. is favored; if neither, this favors 2.

1. Seeing that c was closer to where the accident took place than the others, we take c to be more trustworthy than both: $a \lhd c$ and $b \lhd c$.

2. Because a and b are farther apart than a and c, their viewpoints are likely to be more divergent. Whatever can be observed from widely different perspectives is likely to hold true. Therefore, we will assume $ac \lhd ab$.

3. Because b and c are close together, we add $bc \lhd ac$ as well.

We choose to make no further additions to the relation. In particular, we refrain from making a judgment whether a is more trustworthy than b, or vice versa, or just as trustworthy as b: we consider a and b to be independent. This means it will be consistent with the trust relation to make a choice between which of a and b to trust. If they should happen to contradict each other, our lack of knowledge as to which is more trustworthy then presents us with the option to trust just one of the two.

Note that c is more trustworthy than b in 2. and 3., but that the relationship is not preserved when combined with a ($ac \lhd ab$ holds). Indeed, the following substitution principles are not valid; given $z[y/x] = (z \setminus x) \cup y$,

$$\text{If } x \lhd y \text{ and } x \subset z, \text{ then } z \lhd z[y/x]$$
$$\text{If } x \sim y \text{ and } x \subset z, \text{ then } z \sim z[y/x]\,.$$

2.3 A lattice of trust levels

We know from Lemma 1 that $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$ is a poset. Given the poset it is straightforward to identify the permissible trust attitudes: a trust attitude is permissible if it is an up-set in $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$. Technically, we will represent an attitude by its set of minima, or equivalently, by an antichain in the partial order $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$. We define the set $\mathfrak{T}$ of permissible trust attitudes as follows,

$$\mathfrak{T} = \{\cup\varGamma \mid \varGamma \text{ is an antichain in } (\dot{\mathfrak{S}}, \dot{\trianglelefteq})\}$$

We will use the symbol $\curlywedge$ to denote the attitude that no source unit is trusted, $\cup\emptyset$.

There is a natural relation of strength between permissible trust attitudes. Having a weak trust attitude means trusting only what many sources agree on, or perhaps none; a strong attitude means trusting many sources, or perhaps all. Let $\varGamma$ and $\varDelta$ be antichains in $(\dot{\mathfrak{S}}, \dot{\trianglelefteq})$. Then we define

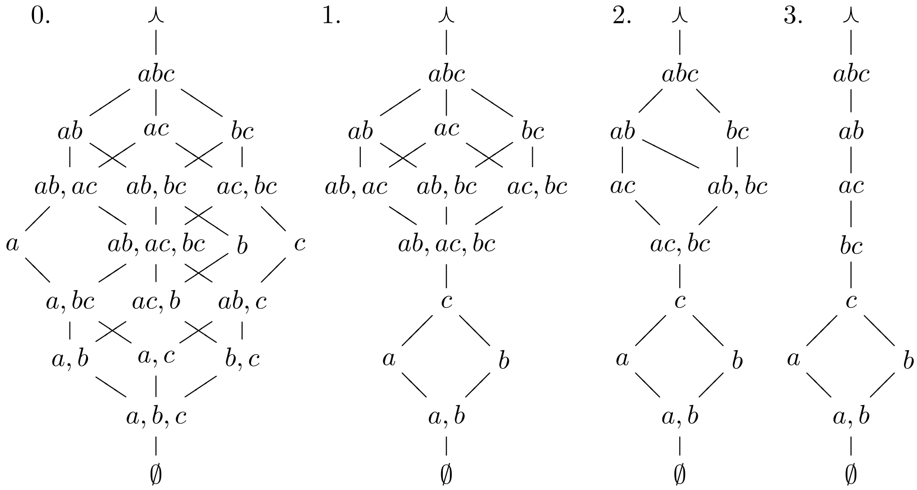
$$\cup\varGamma \leq \cup\varDelta \text{ iff } \uparrow\varDelta \subseteq \uparrow\varGamma.$$

By definition, $\curlywedge$ is $\leq$-maximal in $\mathfrak{T}$. This is natural, as the corresponding attitude of trusting no source unit will always have a maximal degree of reliability. Ordered by $\leq$, the members of $\mathfrak{T}$ form a lattice in which lesser nodes represent stronger trust attitudes. It is natural to talk about the permissible trust attitudes as corresponding to a hierarchy of degrees of trust. We shall hence occasionally refer to $\mathfrak{T}$ as the set of *trust levels*.

In the lattice $(\mathfrak{T}, \leq)$ $A < B$ intuitively means that B is a level of trustworthiness that is genuinely greater than A. Let $\sqcap$ denote meet and $\sqcup$ denote join.

Then $A \sqcup B$ is the weakest trust level that is at least as strong as both A and B; if A and B are not comparable by $\leq$, then it is stronger. $A \sqcap B$ is the strongest trust level that is at least as weak as both A and B.

Example 3 (Lattices for example 2).



The lattice of trust levels makes explicit what the permissible trust attitudes are and how they are related with regard to strength. This can form the basis for choosing, in a given scenario, a *threshold* of trust: a level that is deemed sufficiently trustworthy. Setting a threshold may also be described in terms of *risk*. If $A < B$, then to choose A as the threshold of trust is to take a greater risk with regard to trusting sources than if B is chosen. Determining a threshold of trustworthiness amounts to fixing a "limit" of risk, to draw a line between what is trusted, and not trusted, in the non-relative sense of the word. For example, with a threshold at $A \sqcup B$, if A and B are comparable, risk is limited to what follows from trusting the more trustworthy of the two; if incomparable, then to the greatest degree of risk that represents comparably less risk than both A and B. To say that $A \sqcap B$ lies within the risk limit means that A and B are both considered reliable (i.e., that all source units in A and B provide only true information).

A threshold of trust can be conveniently specified by reference to the source units trusted. Observe that each member of $\dot{\mathfrak{S}}$ is a member of $\mathfrak{T}$. Therefore, any expression using members of $\dot{\mathfrak{S}}$ (i.e., equivalence classes of source units), $\sqcap$ and $\sqcup$ denotes a unique level of trust.

Example 4 (Threshold for example 3). Say that we adopt the attitude to "trust all that ab and ac deliver, as long as it is confirmed by bc" as a threshold. This attitude is expressible as $([ab] \sqcap [ac]) \sqcup [bc]$. Given relations 0., 1., and 2., the attitude amounts to trusting only what a, b, and c agree on, because $(ab \sqcap ac) \sqcup bc = (ab, ac) \sqcup bc = abc$. With the stronger relation 3., it denotes the level ab.

2.4 A tree of *fallbacks* for broken trust

The core of a default conception of relative trust in information sources is the default rule (3) to accept what you are told, unless it is in conflict with what you already know We presently interpret this rule with respect to relative trust. Let us consider a trusting subject that has only permissible trust attitudes. In the non-relative sense of "trust", $⅄$ is always trusted, and a level X is trusted, on condition that every $Y \geq X$ is also trusted, by default.

Now, if trusting at a level X is inconsistent with trusting at a superior level Y, trust at X is broken; X is not trustable. This will obtain whenever information provided at X is inconsistent with antecedent knowledge, or with information accepted at a superior level. The significance of of trusting at X should then be identified with trusting some superior, trustable level; call this the *fallback* of X. The fallback, as the value of a blocked default, is the key notion that allows us to view relative trust as a default attitude.

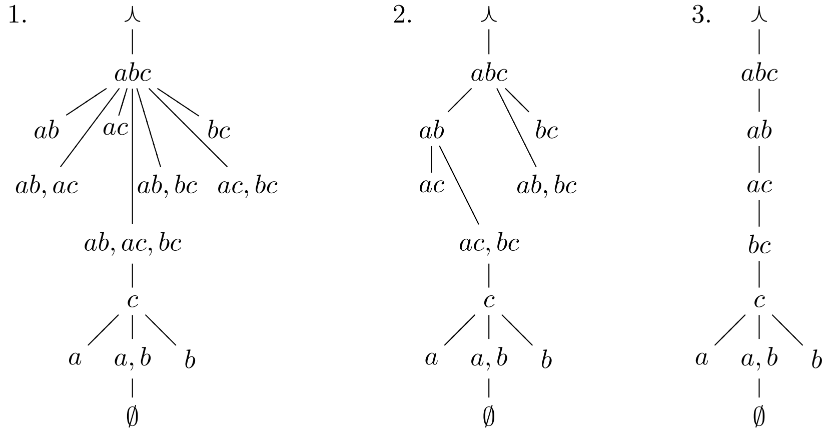
Let X be an element of $\mathfrak{T}$ different from $⅄$, and let Γ be the $\leq$-cover of X. Given that Γ is singleton, we straightforwardly identify $\bigcup \Gamma$ as the appropriate fallback of X. Where not, note that by construction of the lattice, X is a level composed of a set of simpler levels, the members of Γ. That trust is broken at X means some of these levels are not trustable. In this case, the fallback of X should be identified as a level with greater trustworthiness than every Y immediately superior to X. Let the fallback $\mathfrak{f}(X)$ of X be defined as

$$\mathfrak{f}(X) = \mathrm{lub}(\Gamma) \text{ in } (\mathfrak{T}, \leq)\,.$$

The fallback function is undefined for $⅄$; otherwise every node has a unique fallback. $⅄$, representing the trust level of antecedent knowledge, is always the fallback of $[\mathfrak{S}]$. Note that every path from the lattice maximum $⅄$ to a trust level X must go through $\mathfrak{f}(X)$, and that $\mathfrak{f}(X)$ is the $\leq$-minimal node with this property.

The *fallback tree* $(\mathfrak{T}, \prec)$ is defined as the weakest relation such that for all $X \in \mathfrak{T}$, $\mathfrak{f}(X) \prec X$. It is easy to show that the fallback tree is indeed a tree with root $⅄$.

Example 5 (Fallback trees for example 3).



3 Trust in terms of defaults

The aim of this section is to implement the default approach to the information trust model based on a function $\mathbb{I}$ which assigns propositional content to each source in $\mathfrak{S}$. The default interpretation of fallback trees is then encoded into the logic $Æ_{\top}$. Encoding fallback trees in $Æ_{\top}$ will allow us to give precise answers to questions such as, "which trustworthiness levels support a belief in a proposition ϕ?", and "is ϕ entailed by the beliefs of a given degree of trustworthiness?". $Æ_{\top}$ is a natural choice as representation language for default inferences. It allows a simple representation of ordered supernormal defaults theories as well as a natural extension to multi-agent languages.

The basic assignment of information to sources is a mapping from members of $\mathfrak{S}$ to expressions in a formal language. In section 3.1 we use the simple language of propositional logic to this end. However, there is no intrinsic reason for using this language to represent information, and one can easily conceive of using more complex languages for this purpose. Section 3.4 explores possibilities for using multi-agent langauges.

3.1 Information provided by sources

Basically the information interpretation of the trust model assigns formal expressions to each source in $\mathfrak{S}$. The assignment function $\mathbb{I}$ must then be extended to source units (sets of sources) and trust attitudes (sets of source units). To implement this we identify the corresponding operations of *agreement* and *consolidation* of information content. In propositional logic these operations will be implemented simply by means of disjunctions and conjunctions.

Let us denote the informational content of a source a in $\mathfrak{S}$ by $\mathbb{I}_a$, which is a formula of propositional logic. Intuitively, the information $\mathbb{I}_x$ provided by a source unit is defined to be the strongest proposition that every member of the unit supports – the strongest that the members all agree on. If $x = \{a_1, \ldots, a_n\},\ a_i \in \mathfrak{S}$, then $\mathbb{I}_x = \mathbb{I}_{a_1} \vee \cdots \vee \mathbb{I}_{a_n}$. The value of $\mathbb{I}_\emptyset$, on the common understanding of 0-ary disjunctions, will be assumed to be the propositional falsity constant $\bot$. The empty set hence gives a contribution which is always unacceptable.

Define the *consolidated* informational contribution of $x_1, \ldots, x_n \subseteq \mathfrak{S}$ as $\mathbb{I}_X = \mathbb{I}_{x_1} \wedge \cdots \wedge \mathbb{I}_{x_n}$. That is, we define the informational contribution of a set of source units as the strongest consequence that would follow from taking each unit as a source of evidence. Observe in particular that $\mathbb{I}_{[\emptyset]}$ will always be $\bot$. By convention $\mathbb{I}_{\curlywedge}$ is $\top$.

3.2 Intermezzo: The Logic $Æ_{\top}$

$Æ_{\top}$ is an "Only knowing" logic, generalizing the pioneering system of Levesque [7] with language constructs for the representation of various degrees of confidence for a doxastic subject.

The object language of $Æ_{\top}$ extends the language of propositional logic by the addition of modal operators: $\Box$ (necessity) and modalities B_k (belief) and

C_k (co-belief) for each k in a finite index set I. The index set represents the distinct degrees of confidence and comes along with a partial order which gives the indices relative strength. $\mathsf{b}_k\,\varphi$ is defined as $\neg\,\mathsf{B}_k\,\neg\varphi$ and denotes that φ is compatible with belief at degree of confidence k.

A formula φ is *completely modalized* if every occurrence of a propositional letter occurs within the scope of a modal operator and *purely Boolean* if it contains no occurrences of modal operators. The "all I know at k" expression $\mathsf{O}_k\,\varphi$ abbreviates $\mathsf{B}_k\,\varphi \wedge \mathsf{C}_k\,\neg\varphi$, meaning that *precisely* φ is believed with degree of confidence k. A formula of the form $\bigwedge_{k\in I} \mathsf{O}_k\,\varphi_k$ is called an O_I*-block*. If each φ_k is purely Boolean, the O_I-block is said to be *prime*.

$\text{Æ}_\top$ is a special instance of the system Æ_ρ introduced in [8] and further analyzed and motivated in [12]; the references contain in particular an axiomatization, a formal semantics and proofs of soundness, completeness and the finite model property. A particularly strong property of $\text{Æ}_\top$ is the Modal Reduction Theorem: for each O_I-block φ^I and for some $m \geq 0$, there are prime O_I-blocks $\psi_1^I, \ldots, \psi_m^I$ such that $\vdash \varphi^I \equiv (\psi_1^I \vee \cdots \vee \psi_m^I)$.[2]

A prime O_I-block determines the belief state of the agent in a unique and transparent way; if such a formula is satisfiable, it has essentially only one model. A non-prime O_I-block only implicitly defines the belief state and has in general a number of different models. The Modal Reduction Theorem relates an implicit belief representation to an explicit representation by a provable equivalence. To determine whether $m > 0$ in the statement of the theorem is Σ_2^p-hard.

If there is only one degree of confidence, $\text{Æ}_\top$ is equivalent to Levesque's system of only knowing, for which there is a direct correspondence between a stable expansion in autoepistemic logic and a prime formula $\mathsf{O}\,\varphi$. A prime O_I-block is a natural generalization of the notion of stable expansion to a hierarchical collection of expansions.

3.3 Encoding the fallback tree as defaults in $\text{Æ}_\top$

We now describe how to use a fallback tree to extract information, both between contributions of the sources, which may be more or less mutually compatible, and between these contributions and a set of antecedently given information.

To facilitate the discussion let us say that a fallback tree is *information labelled* if each node X in the tree is labelled with $\mathbb{I}_X$. The labels express the information contribution attached to the trust level X.

We will assume that a *knowledge base*, denoted κ, is given with unconditional trustworthiness. Informally, say that (precisely) κ, a formula of propositional logic, is believed with full conviction. The notion of trustworthiness is directly relevant to the notions of confidence and belief, as is clear by the simple observation that information stemming from highly trustworthy sets of sources will be considered reliable with a greater degree of confidence than that which is

[2] In the sequel $\vdash$ denotes the provability relation of $\text{Æ}_\top$ (which extends the provability relation of classical logic).

provided by less trustworthy sources. Following the default interpretation formulated in principle (3), we can define a simple procedure which reveals what information may reliably be said to be supported at each level of trustworthiness. Define the following formula by induction over the fallback tree.

$$\begin{aligned} \beta_{\Yup} &= \kappa \\ \beta_X &= \begin{cases} \beta_{\mathfrak{f}(X)} \wedge \mathbb{I}_X & \text{if } \beta_{\mathfrak{f}(X)} \wedge \mathbb{I}_X \text{ is PL-consistent,} \\ \beta_{\mathfrak{f}(X)} & \text{otherwise.} \end{cases} \end{aligned}$$

Then β_X denotes what a rational agent should believe at a degree of confidence corresponding to the trust attitude X.

The modal logic $\text{Æ}_\top$ is suitable for the representation of fallback trees and the associated default principle. In the encoding we use the set of trust levels $\mathfrak{T}$ as the index set which individuates modalities in the language of $\text{Æ}_\top$. Let $(\mathfrak{T}, \prec)$ be the fallback tree and $\prec^*$ be the reflexive, transitive closure of $\prec$. For $X \in \mathfrak{T}$ we define

$$\delta_X = \mathsf{b}_X \, \mathbb{I}_X \supset \mathbb{I}_X \, .$$

Note that δ_X is equivalent to $\neg \, \mathbb{I}_X \supset \mathsf{B}_X \neg \, \mathbb{I}_X$, i.e., should φ be false, the subject will believe that it is. We will refer formulae of this form as *default conditionals* when they occur within a modal O-context, since the conditional then has the force of formalizing the property corresponding to the statement "the proposition $\mathbb{I}_X$ holds by default".

The default interpretation of the default structure is formalized by the following encoding:

$$\begin{aligned} [\![\mathfrak{T}, \prec, \kappa]\!]_{\Yup} &= \mathsf{O}_{\Yup} \, \kappa \\ [\![\mathfrak{T}, \prec, \kappa]\!]_X &= \mathsf{O}_X(\kappa \wedge \textstyle\bigwedge_{Y \prec^* X} \delta_Y) \\ [\![\mathfrak{T}, \prec, \kappa]\!] &= \textstyle\bigwedge_{X \in \mathfrak{T}} [\![\mathfrak{T}, \prec, \kappa]\!]_X \end{aligned}$$

The encoding is structurally similar to the encoding of ordered default theories into $\text{Æ}_\top$ in [4].

Theorem 1. $\vdash [\![\mathfrak{T}, \prec, \kappa]\!] \equiv \bigwedge_{X \in \mathfrak{T}} \mathsf{O}_X \, \beta_X$.

Proof. The proof uses simple properties from the model theory of $\text{Æ}_\top$, cf. [12]. In an $\text{Æ}_\top$ model M all points agree on the truth value of every completely modalized formula. We will hence use the notation $M \models \varphi$ whenever a completely modalized φ is satisfied at some point in M. We use the following two facts in the proof. Let M satisfy $\mathsf{O}_X \, \varphi$ for an index X.

1. If M satisfies $\mathsf{O}_X \, \psi$, then $\varphi \equiv \psi$ is true at every point in M.
2. If φ and ψ are purely Boolean, M satisfies $\mathsf{b}_X \, \psi$ iff $\varphi \not\vdash \neg\psi$.

We show, by induction on X, the more general result that for any $Z \in \mathfrak{T}$

$$\vdash \textstyle\bigwedge_{X \prec^* Z} [\![\mathfrak{T}, \prec, \kappa]\!]_X \equiv \textstyle\bigwedge_{X \prec^* Z} \mathsf{O}_X \, \beta_X \quad .$$

The base case is trivial. For the induction step, it is sufficient to show that $M \models [\![\mathfrak{T}, \prec, \kappa]\!]_X \equiv \mathsf{O}_X\, \beta_X$ for any Æ$_\top$-model satisfying both $[\![\mathfrak{T}, \prec, \kappa]\!]_{\mathfrak{f}(X)}$ and $\mathsf{O}_{\mathfrak{f}(X)}\, \beta_{\mathfrak{f}(X)}$. By 1, every such model M satisfies

$$M \models (\kappa \wedge \bigwedge\nolimits_{Y \prec^* \mathfrak{f}(X)} \delta_Y) \equiv \beta_{\mathfrak{f}(X)} \ \ .$$

Thus $M \models [\![\mathfrak{T}, \prec, \kappa]\!]_X \equiv \mathsf{O}_X(\beta_{\mathfrak{f}(X)} \wedge \delta_X)$. It only remains to show

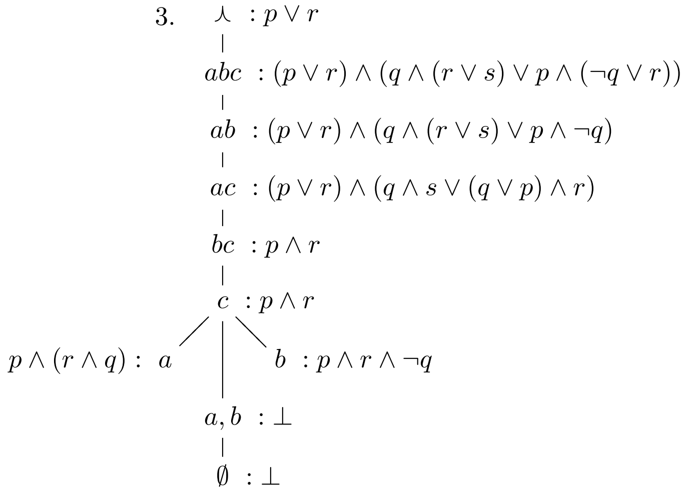
$$M \models \mathsf{O}_X(\beta_{\mathfrak{f}(X)} \wedge (\mathsf{b}_X\, \mathbb{I}_X \supset \mathbb{I}_X)) \equiv \mathsf{O}_X\, \beta_X \ \ .$$

But since $M \models \mathsf{O}_{\mathfrak{f}(X)}\, \beta_{\mathfrak{f}(X)}$, it follows directly from the definition of β_X and 2 that $M \models \mathsf{b}_X\, \mathbb{I}_X$ iff $\beta_{\mathfrak{f}(X)} \not\vdash \neg \mathbb{I}_X$, and we are done. □

The theorem shows that the encoding of a node X and its information content can be reduced to the $\mathsf{O}_{\mathfrak{T}}$-block $\bigwedge_{X \in \mathfrak{T}} \mathsf{O}_X\, \beta_X$ within the logic itself, where at each node X in the tree the formula β_X is the proposition that the rational agent will entertain at this level of trust.

Example 6 (Example 5, with information). The witnesses a, b, and c are interviewed for their accounts of the accident scenario. We assign content to propositional variables as follows: p = The green car was veering; q = There was a cat in the road; r = The red car was veering; s = The red car was speeding. The following figure records the witnesses' statements (left), and the resulting post-evaluation propositions at each trust level decorate the fallback tree (3.).[3]

$a \ : \ q \wedge (r \vee s)$
$b \ : \ p \wedge \neg q$
$c \ : \ p \wedge r$


3.
⅄ : $p \vee r$
abc : $(p \vee r) \wedge (q \wedge (r \vee s) \vee p \wedge (\neg q \vee r))$
ab : $(p \vee r) \wedge (q \wedge (r \vee s) \vee p \wedge \neg q)$
ac : $(p \vee r) \wedge (q \wedge s \vee (q \vee p) \wedge r)$
bc : $p \wedge r$
c : $p \wedge r$
$p \wedge (r \wedge q)$: a
b : $p \wedge r \wedge \neg q$
a, b : $\bot$
$\emptyset$: $\bot$

Noteworthy features:

- a and b may not both be fully trusted, but choosing either is consistent.
- The proposition s, which figures as a disjunct in a's account, is eliminated from the node bc onwards.

[3] Formulae computed using *The Logics Workbench*, `http://www.lwb.unibe.ch/`.

- For nodes a, b and $\emptyset$, the value $\bot$ is displayed to emphasize their inconsistency. These nodes will actually take values from the consistent fallback node c, i.e., $p \wedge r$.

3.4 From information sources to doxastic agents

There is no intrinsic reason to use the language of propositional logic to represent the information delivered by sources. This section addresses the use of multimodal languages for this purpose. The expressive power of such languages is needed in cases where the sources deliver information about agents; typically, about what the agents believe. To generalize the approach of section 3.3 we need to extend the language of $\text{Æ}_\top$ such that it extends the information representation language.

The logic $\text{Æ}_\top$ has been extended to a multi-modal language. An interesting proof-theoretical property of this extension of $\text{Æ}_\top$ is that it has a sequent calculus formulation which admits constructive cut-elimination and hence cut-free proofs; this is proved in [11] for a multi-agent language in which the beliefs of each subject are represented relative to different degrees of confidence. A Kripke semantics for the logic has been presented in [13].[4]

Let us assume that the modalites in the multi-agent language is defined by a collection $I_0, \ldots, I_m$ of index sets, one for each agent. The indices in each index set are partially ordered, while two indices in different index sets are unrelated.

The notion of an O_I-block transfers to the multimodal langage: An O_{I_j}-block is a formula $\bigwedge_{k \in I_j} \mathsf{O}_k\, \varphi_k$. If each formula φ_k is I_j-objective, i.e. all occurrences of a I_j-modality occurs within the scope of a modality which belongs to another agent, the O_{I_j}-block is prime. An O_I-block can now be defined as a conjunction of O_{I_j}-blocks, one for each agent. Given these concepts the Modal Reduction Theorem transfers to the multi-modal logic.

Let us first assume that the sources deliver information about the beliefs of agents $\alpha_1, \ldots, \alpha_m$ without being agents themselves, i.e. they do not deliver information about other sources, or about themselves, or about the observer who collects the information. Assume also that the beliefs of these agents are represented in the multi-modal system K45$_m$, i.e. a sublanguage of multi-modal $\text{Æ}_\top$, so that the $\mathbb{I}$ function now delivers K45$_m$ formulae.

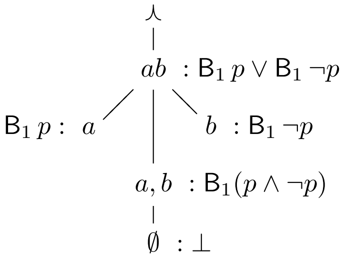
The index sets for the multi-modal $\text{Æ}_\top$-representation are $\mathfrak{T}, \{\alpha_1\}, \ldots, \{\alpha_m\}$. The assumption that no α_i are sources implies that we can use the same simple functions for agreement and consolidation as introduced for propositional logic in section 3.1. It is now straightforward to establish Theorem 1 for the language at hand.

Example 7 (Modal information). A simple case in which sources provide formulae in a modal language. Let the trustworthiness relation be given as in example 1, relation 3. Let the knowledge base be empty, and assign information to sources

[4] The semantics has been given for a multi-agent language without confidence levels. An extension to the languge addressed in this section is straightforward.

as below (left). The fallback tree shows the outcome of evaluation (right). Here, trusting what a and b agree on (source unit ab) implies accepting that agent 1 has a full belief regarding p. Trusting both sources (node a, b) implies accepting that 1 is inconsistent.

$$\begin{array}{lcl} a & : & \mathsf{B}_1\, p \\ b & : & \mathsf{B}_1 \neg p \end{array}$$


⅄
ab : $\mathsf{B}_1\, p \vee \mathsf{B}_1 \neg p$
$\mathsf{B}_1\, p$: a
b : $\mathsf{B}_1 \neg p$
a, b : $\mathsf{B}_1 (p \wedge \neg p)$
$\emptyset$: $\bot$

If the information sources are themselves agents, the situation is at once much more complex, and we propose this to other researchers in the community as an interesting and challenging application of multi-modal logics. One problem is that we can no longer implement agreement and consolidation by means of simple Boolean operations. In some cases we may use the notion of "group belief" for agreement and "distributed belief" for consolidation (see e.g. [5]).

However, we can also use the full expressive power of multi-modal $\text{Æ}_{\top}$ to specify very complex formulae delivered by each agent, in which case these operators are no longer sufficient for this purpose. We plan to address this in a follow-up paper.

4 Related work

The present account of trustworthiness generalizes and clarifies the approach introduced by John Cantwell [1]. Our approach improves on Cantwell's by making a clear separation between the notion of trustworthiness on the one hand, and information and belief on the other, which allows for the notion of trustworthiness level to be separated from a given model. Furthermore, the present theory gives informative results for various weak kinds of trustworthiness relations that yield vacuous output on Cantwell's approach.[5]

In this paper, no attempt has been made to give a general account of the basic non-relative notion of trust; for this, see Jones [6]. We intend to apply the present theory of relative trustworthiness to Jones' analysis of trust in a forthcoming paper. We also wish to explore the complex subjects of construction and revision of trustworthiness relations in the future.

[5] Cantwell incorporates his theory of trustworthiness into a theory of *belief revision*. This is an application that we have not gone into.

Bibliography

[1] John Cantwell. Resolving conflicting information. *Journal of Logic, Language, and Information*, 7:191–220, 1998.

[2] John Cantwell. *Non-Linear Belief Revision*. Doctoral dissertation, Uppsala University, Uppsala, 2000.

[3] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2 edition, 2002.

[4] Iselin Engan, Tore Langholm, Espen H. Lian, and Arild Waaler. Default reasoning with preference within only knowing logic. Submitted for publication.

[5] Ronald Fagin, Joseph Y. Halpern, Yoram Moses, and Moshe Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, Mass., 1995.

[6] Andrew J. I. Jones. On the concept of trust. *Decision Support Systems*, 33:225–232, 2002.

[7] Hector J. Levesque. All I know: A study in autoepistemic logic. *Artificial Intelligence*, 42:263–309, 1990.

[8] E. H. Lian, T. Langholm, and A Waaler. Only knowing with confidence levels: Reductions and complexity. In *JELIA 2004*, Logics in Artificial Intelligence, 9th European Conference, pages 500–512, Lisbon, Portugal, 2004.

[9] K. Segerberg. Some modal reduction theorems in autoepistemic logic. *Uppsala Prints and Preprints in Philosophy*, 1995.

[10] Arild Waaler. *Logical Studies in Complementary Weak S5*. Doctoral thesis, University of Oslo, Oslo, 1994.

[11] Arild Waaler. Consistency proofs for systems of multi-agent only knowing. *Advances in Modal Logic*, 2005.

[12] Arild Waaler, Johan W. Klüwer, Tore Langholm, and Espen H. Lian. Only knowing with degrees of confidence. Submitted for publication, 2005.

[13] Arild Waaler and Bjørnar Solhaug. Semantics for multi-agent only knowing (extended abstract). Submitted for publication, 2005.