EPiC Series in Computer Science
Volume 36, 2015, Pages 215–228
GCAI 2015. Global Conference on Artificial Intelligence

Trust, Belief and Honesty*

David Pearce[1] and Levan Uridia[2]

[1] Universidad Politecnica de Madrid, david.pearce@upm.es
[2] Razmadze Institute of Mathematics, uridia@ia.urjc.es

Abstract

A well-known, early formal treatment of trust as a modality can be found in the work of Liau [9]. Recently in [4] simplified versions of Liau's logic were studied. In this paper we examine two extensions of the modal logics of trust from [4], formed by adding further axioms. The idea is to interconnect the trust modality with the individual belief modalities of agents. In the first of our logics we capture the idea that if an agent i trusts an agent j with respect to a statement p then i believes that j does not disbelieve p; while in the second logic if i trusts j concerning p, then i believes that j believes p. In this way we can explicate a type of trust that is linked to honesty or sincerity. These quite intuitive extensions of the logic of trust help to solve some unintuitive consequences that arise when the semantics of trust and belief are independent. As a technical result we prove the soundness and completeness of these logics.

1 Introduction

The notion of *trust* is widely recognized to be a key concept of social intelligence and an important foundational concept for socio-cognitive technical systems. It has been investigated in various disciplines including logic, philosophy, economics, sociology, cognitive science and computer science. A recent and significant work on the logic of trust (and reputation) [8] also provides a useful characterisation of various trust concepts as well as an extensive bibliography.

There are several kinds of trust in common usage. Dictionary definitions of trust refer to a belief in or reliance on a certain property of an individual or thing, for example the "firm belief in the reliability or truth or strength of a person or thing" (Oxford Pocket Dictionary), or the "assured reliance on the character, ability, strength, or truth of someone or something" (Merriam-Webster online). The Cambridge dictionary online does not mention truth but does pick out the important property of *honesty*: "to believe that someone is good and honest and will not harm you, or that something is safe and reliable". These different traits of trust are partly independent. Trusting someone (about a judgement) with respect to their honesty is not the same as trusting them for their reliability or credibility. Suppose I want to buy a used electrical appliance at a flea market where there are no facilities to check it over. I may trust the trader that he is not cheating me when he says that he believes the appliance is working, yet I may on

*The first author was partially supported by the project MERLOT, TIN 2013-42149-P. The second author was partially supported by Rustaveli Science Foundation project number 52/05. Both authors were partly supported by SINTELNET, FP7-ICT-2009-C Project No. 286370. An earlier version of this paper was delivered at LAMAS 2015 under the title 'An Honest View of Trust'. We are grateful to the many suggestions for improvements that we have received from anonymous referees.

inspection find some evidence of damage that suggests to me a possible malfunction that leads me to reject the item. There are also recent approaches to the study of trust from a dynamic epistemic logic perspective [1], [14]. In this paper we examine a logical model of trust in the tradition of those accounts that [8] regards as representing an *informational* view of trust. In this tradition, trust is viewed as a propositional attitude that is usually linked to beliefs of a weaker or stronger kind, for example that if one agent trusts another agent (with respect to a judgement p), then this may entail that the former agent supposes that the other agent believes p, or perhaps even that p is actually true.

We will focus on the interrelation of trust and belief for a finite sets of agents. Our point of departure is the logical model of trust proposed by Churn-Jung Liau in [9]. Although Liau's work belongs to the tradition of informational models of trust, in this respect we will simplify Liau's approach in that we retain only two types of modalities: $\Box_i$ as a belief operator of agent i and $T_{i,j}\varphi$ which stands for the trust of agent i in agent j with respect to a proposition φ. We do not make use of the additional operator I of information passing. In this paper, the idea that when a trust relation is formed it may be based on the passing of information φ from one agent to another *will be left implicit throughout*. So the logic $\mathbf{L1}$ we introduce is especially designed to talk about belief and trust and their interrelation.

Another difference with respect to [9] is that instead of adopting $\mathbf{KD45}$ as a doxastic logic of belief, we use the logic $\mathbf{KS}$ first introduced in the 1970s by Krister Segerberg as a modal logic of *some other time* [16]. $\mathbf{KS}$ was later rediscovered as a modal logic of *inequality* [6], [13], and recently it was adopted as a doxastic logic concurrent to the classical system $\mathbf{KD45}$ [2]. In this paper we restrict attention to those agents whose doxastic properties are formalized in the modal logic $\mathbf{KS}$. From a technical point of view we could also use a stronger doxastic base logic. For instance, similar formal results would hold if we would add axiom $\mathbf{D}$ to our base logic. Or we could even work in $\mathbf{KD45}$. Our preference for $\mathbf{KS}$ depends on various properties of it that we have explored in previous work. In particular, there are various reasoning contexts, important in AI, where the $\mathbf{KD45}$ seems too strong as a basis for beliefs. For example, in the context of nonmonotonic reasoning, the standard extension of $\mathbf{KD45}$ would be autoepistemic logic, which we may fairly judge to be a logic of knowledge. In multi-agent settings, however, it is important to be able to distinguish between knowledge and belief, eg. negative introspection, corresponding to axiom $\mathbf{5}$, is questionable in several doxastic contexts and definitely false in some, such as *common belief*. We have discussed some of the good features of $\mathbf{KS}$ (and its variants) as a basis for doxastic logic in some previous papers. Eg. in [10] we apply the logic $\mathbf{KSD}$ to the study of minimal belief; in [11] we apply the extension $\mathbf{K4}$ to the study of common belief.

In Liau's base logic $\mathbf{BIT}$, if an agent i trusts an agent j with respect to a proposition φ and moreover believes that the information that φ has been received from j, then i believes φ. This is an axiom of the system. On the other hand, Liau's base logic does not entail that i believes that j believes φ, even if i trusts j about φ and believes that j has informed him of φ. This last condition is considered in [9] but as part of an extended trust concept called *cautious trust*. This concept also includes the reliability feature of trust that the trusted agent is not only truthful but also truthlike. Expressed in our notation, the extended concept satisfies: $T_{i,j}\varphi \to \Box_i(\Box_j\varphi \to \varphi)$. By standard modal logic, it follows that once i trusts j, he is thought to be reliable if he is thought to be honest.[1]

Rather than conflate honesty and reliability, our aim is build a logic of trust that, while following the semantical treatment of the trust operator given in [9], allows us to consider different forms of honesty separately and to incorporate these already into the base logic. In particular, we want to consider a system that contains

$$\vdash T_{i,j}p \to \Box_i\Box_j p \tag{1}$$

as an axiom. In Robert Demolombe's formal treatment of trust [3], this corresponds to what he calls

[1] In fact in $\mathbf{BIT}$ cautious trust (together with information exchange) already entails directly that the trusting agent believes φ.

sincerity.[2] Quite rightly, this is distinguished from what he terms *credibility*, which entails that

$$\vdash T_{i,j}p \to \Box_i(\Box_j p \to p)$$

ie., precisely the condition just mentioned as part of Liau's cautious trust concept. We shall also consider an intuitively weaker concept of honesty or sincerity expressed by the condition

$$\vdash T_{i,j}p \to \Box_i \Diamond_j p. \quad (2)$$

Here $\Diamond$ is simply the dual operator to $\Box$, i.e., $\Diamond_j p \equiv \neg\Box_j\neg p$. So this represents an alternative form of sincerity that amounts to the agent i believing that he is not being cheated or deceived by j in the sense that j actually believes the contrary of p. (1) and (2) are the basic forms of honesty or sincerity that we think deserve to be studied alongside the trust relation. We could imagine a still weaker condition like $\neg\Box_i\Box_j\neg p$, i.e., i does not believe he is being deceived. But while this should certainly belong to the idea of trust as belief in sincerity, it seems too weak to work on its own as a basis for trust. So our basic system will include either (1) or (2) as axioms.[3]

The remainder of the paper is organised as follows. In Section 2 we recall the logic **KS**, providing some basic definitions and some known facts. In Section 3 we discuss Liau's approach and present an example which helps to motivate our work. In Section 4 we introduce the logics **L1** and **L2** and study their semantics on structures with neighbourhood functions. As a main result we prove completeness of the logics with respect to the given semantics. In Section 5 we draw conclusions and discuss future work.

2 The Modal Logic KS

In 1976 Krister Segerberg [16] formulated a modal logic **KS** in which the diamond modality $\Diamond$ is interpreted as "somewhere else". In this section we define the system **KS** and its semantics.

2.1 Syntax

The normal modal logic **KS** is defined in a standard modal language signature with infinite set of propositional letters $Prop = \{p, q, r, ..\}$ and connectives $\wedge, \Box, \neg$. Formulas are built up inductively according to the grammar:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid \Box\phi,$$

where p ranges over propositional symbols in $Prop$. The logical symbols '$\top$' and '$\bot$', and the additional connectives such as '$\vee$', '$\to$' and '$\leftrightarrow$' and the dual modality '$\Diamond$', are defined as usual, i.e., $\top := p \vee \neg p$ for some atomic proposition p; $\bot := \neg\top$; $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$; $\phi \to \psi := \neg\phi \vee \psi$;$\phi \leftrightarrow \psi := (\phi \to \psi) \wedge (\psi \to \phi)$; and $\Diamond\phi := \neg\Box\neg\phi$.

Axioms are all classical tautologies plus three axioms containing modal operators. Namely:

(K) $\Box(p \to q) \to \Box p \to \Box q$;

$(w4)$ $\Box p \wedge p \to \Box\Box p$;

(B) $p \to \Box\Diamond p$; where $\Diamond p \equiv \neg\Box\neg p$.

[2] We ignore here the aspect of information exchange and a distinction that Demolombe makes between belief and strong belief.

[3] Once communication is explicitly taken into account, various types of sincerity (and dishonesty) can be distinguished. For a formal treatment, see [15].

The directed graph represents a structure with irreflexive points coloured grey and reflexive points uncolored. As we can see yRx and xRy but not yRy, which contradicts transitivity but not weak transitivity.

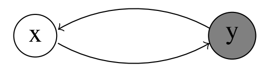

Figure 1:

Rules of inference are: Modus-ponens, Substitution and Necessitation.

Observe that doxastic interpretation of the axiom (B) states that *If* p *is true then agent believes that it is not the case that he believes negation of* p. It is not difficult to show that if we add axion $\Box p \to p$ to $\mathbf{KS}$ we will get $\mathbf{S5}$. Following Smullyan [17] this means that if a $\mathbf{KS}$ - reasoner is *accurate* (never believes any false proposition) then his beliefs coincide with his knowledge. A brief study of the axiom $w4$ and in particular the logic $\mathbf{KS}$ can be found in [5].

2.2 Possible Worlds Semantics

We begin by recalling the main semantical concepts such as (possible worlds) model, satisfaction and the validity of modal formulas. These definitions are standard and can be found in any modal logic book. A (possible world) *model* is a pair $\mathcal{M} = (\mathfrak{M}, V)$, where $\mathfrak{M} = (W, R)$ is a *frame*, with W a non-empty set of points or 'worlds' and R a binary relation on W; while $V : W \to 2^W$ is the valuation function. An interpretation of formulas is given by means of a binary relation, '$\Vdash$', between pointed models and formulas. A pointed model is a pair $\langle \mathcal{M}, w \rangle$, where $\mathcal{M} = (W, R, V)$ is a model and w a world from W. The satisfaction relation is defined inductively on the structure of formulas φ as:

- $\langle \mathcal{M}, w \rangle \Vdash p$ iff $w \in V(p)$;
- $\langle \mathcal{M}, w \rangle \Vdash \neg\psi$ iff $\langle \mathcal{M}, w \rangle \nVdash \psi$;
- $\langle \mathcal{M}, w \rangle \Vdash \psi \wedge \chi$ iff $\langle \mathcal{M}, w \rangle \Vdash \psi$ and $\langle \mathcal{M}, w \rangle \Vdash \chi$;
- $\langle \mathcal{M}, w \rangle \Vdash \Box\psi$ iff for all $v \in W$ with $(w, v) \in R$, $\langle \mathcal{M}, v \rangle \Vdash \psi$.

A formula ϕ is said to be *true* at w in $\mathcal{M}$ if $\langle \mathcal{M}, w \rangle \Vdash \phi$; ϕ is *satisfiable* if there is a pointed model $\langle \mathcal{M}, w \rangle$ at which it is true; ϕ is *valid in* $\mathcal{M}$ (written '$\mathcal{M} \Vdash \phi$') if $\langle \mathcal{M}, w \rangle \Vdash \phi$ for all w in $\mathcal{M}$.

The semantics for the modal logic $\mathbf{KS}$ is provided by weakly-transitive and symmetric frames. Below we give the definition of weakly-transitive relation.

Definition 1. *We will say that a relation* $R \subseteq W \times W$ *is weakly-transitive if* $(\forall x, y, z) \quad (xRy \wedge yRz \wedge x \neq z \Rightarrow xRz)$.

Of course every transitive relation is weakly transitive as well. Moreover it is not difficult to notice that weakly transitive relations differ from transitive ones just by the occurrence of irreflexive points inside clusters. Figure 1 represents a frame which underlines the main difference between weakly transitive and transitive frames.

In the study of modal logic the class of rooted frames plays a central role. Recall that a frame $\mathfrak{M} = (W, R)$ is rooted if it contains a point $w \in W$, which can see all other points in W. That is $R(w) \supseteq W \setminus \{w\}$, where $R(w)$ is the set of all successors of w. The class of all rooted, weakly-transitive and symmetric frames can be characterized by the property which we call weak - cluster.

Definition 2. *We will say that a relation* $R \subseteq W \times W$ *is weak-cluster if* $(\forall x, y)(x \neq y \Rightarrow xRy)$.

It is easy to see that every weak-cluster is a universal relation where we allow irreflexive points. The following proposition links weak-clusters with rooted, weakly-transitive, symmetric frames.

Proposition 3. *A frame* $\mathfrak{M} = (W, R)$ *is rooted, weakly-transitive and symmetric iff it is weak-cluster.*

Proof. It is immediate that every weak-cluster is a rooted, weakly-transitive and symmetric frame. For the other direction let $\mathfrak{M} = (W, R)$ be a rooted, weakly-transitive and symmetric frame. Let $w \in W$ be the root. Take two arbitrary distinct points $x, y \in W$. As w is the root, we have: wRx and wRy. Because of symmetry we get xRw. Now as R is weakly-transitive, from $xRw \wedge wRy$ and $x \neq y$ we get xRy. Hence R is a weak-cluster. □

So far we have defined the modal logic $\mathbf{KS}$ syntactically and we gave the definition of weak-cluster relation. The following theorem links these two notions:

Theorem 4. *[5] The modal logic* $\mathbf{KS}$ *is sound and complete w.r.t. the class of all* ***finite, irreflexive*** *weak-cluster relations.*

Mainly because of the theorem 4 the modal logic $\mathbf{KS}$ is called the modal logic of inequality. As the reader can easily check the interpretation of box in irreflexive weak-clusters boils down to the following: $w \Vdash \Box\phi$ iff $(\forall v)(w \neq v \Rightarrow v \Vdash \phi)$.

3 Some (brief) State of the Art

We first provide a short overview of the state of research on trust. Two main directions are involved. Let us call them *practical* and *foundational*. The practical direction is related with the rise of virtual communities such as online chat rooms, electronic markets, and virtual multiplayer game worlds. These new spaces of interaction have challenged our accumulated wisdom on how agent interactions can occur. Many aspects of these virtual communities have been extensively researched and trust and reputation are widely described as central to effective interactions between agents [12].

The other, foundational direction, is, with some exceptions, even more recent. With an increasing number of applied works, the need for a foundational background has grown and is being addressed at the interface of philosophy and computer science. Many works on trust and reputation have adopted a quantitative representation of these concepts. Trust and reputation are commonly simplified to a numerical representation losing important properties of these concepts. Steps towards investigating foundational properties (on the level of mathematical logic) as a bridge from philosophical foundations towards the logical foundations of these concepts were made in [8]. Even earlier one of the pioneering works on trust and information acquisition [9] succeeded in establishing interesting links between logical systems and some existing models in different branches of applied computer science. Both of the works mentioned use modal logic as a background formalism, although they use logic in different ways. In [8] the approach is more *external*: several different notions of trust are defined in terms of 'simpler' notions (modalities) such as knowledge, belief, choice etc. This allows one to distinguish on a formal level different kinds of trust met in the philosophical literature. In [9] the approach is more *internal*, i.e. the trust modality is taken as a primitive notion and properties of trust are studied in terms of trust, belief and information acquisition. In this paper we follow this second approach.

3.1 Liau's Approach

Let us recall Liau's approach. The language of his logic **BIT** is a standard propositional language with several modal operators of different types: information acquisition operators, trust operators and belief operators $I_{i,j}$, $T_{i,j}$ and $\Box_i$ respectively for each pair of distinct agents $i, j \in A$. Each $\Box_i$ modality satisfies the **KD45** axioms and each $I_{i,j}$ modality satisfies the axioms of **KD**. i.e. we have $\Box_i(p \to q) \to (\Box_i p \to \Box_i q)$, $\Box_i p \to \Box_i \Box_i p$, $\Box_i p \to \Diamond_i p$, $\Diamond_i p \to \Box_i \Diamond_i p$ for each index i and $I_{i,j}(p \to q) \to (I_{i,j} p \to I_{i,j} q)$ and $I_{i,j} p \to \neg I_{i,j} \neg p$ for each pair of indexes i, j. For the trust modalities $T_{i,j}$ there are two main axiom schemes reflecting the following intuitive ideas:

1) agent i trusts agent j if and only if agent i believes that he trusts agent j. This is represented by the axiom $T_{i,j} p \leftrightarrow \Box_i T_{i,j} p$.

2) If agent i believes that the information p came from agent j and he trusts agent j about p then he believes the information p. This is given by: $\Box_i I_{i,j} p \wedge T_{i,j} p \to \Box_i p$.

The second axiom provides a relation between trust and belief, although this connection is rather between information acquisition operator and beliefs of agents. If we ignore the information acquisition operator then the connection vanishes and trust and belief become totally independent from each other. This is also reflected on semantical level.

The semantics of the logic **BIT** is provided by frames $(W, R_i, S_{i,j}, u_{i,j})$ where W is a set of possible worlds, each R_i is transitive, serial and euclidean relation (**KD45**-relation), each $S_{i,j}$ is transitive and serial relation (**KD**-relation) and each $u_{i,j} : W \to \mathcal{PP}(W)$ is a neighbourhood function that assigns each world a family of sets of worlds. The main interesting novelty of the approach is to provide semantics for the trust operator via neighbourhood functions.

Definition 5. *A satisfaction of a formula in a given model $\mathcal{M} = (W, R_i, S_{i,j}, u_{i,j}, V)$ and a point $w \in W$ is defined inductively as follows:*
$w \Vdash p$ *iff* $w \in V(p)$;
$w \Vdash \neg\alpha$ *iff* $w \nVdash \alpha$;
$w \Vdash \alpha \wedge \beta$ *iff* $w \Vdash \alpha$ *and* $w \Vdash \beta$*;*
$w \Vdash \Box_i \alpha$ *iff* $(\forall w')(w R_i w' \Rightarrow w' \Vdash \alpha)$*;*
$w \Vdash I_{i,j} \alpha$ *iff* $(\forall w' \in W)(w S_{i,j} w' \Rightarrow w' \Vdash \alpha)$*;*
$w \Vdash T_{i,j} \alpha$ *iff* $|\alpha| \in u_{i,j}(w)$. *Here* $|\alpha|$ *denotes the set* $\{v \in W \mid v \Vdash \alpha\}$.

The axioms $T_{i,j} p \leftrightarrow \Box_i T_{i,j} p$ and $\Box_i I_{i,j} p \wedge T_{i,j} p \to \Box_i p$ define the following two properties,

$$u_{i,j}(w) = \bigcap_{v \in R_i(w)} u_{i,j}(v); \tag{3}$$

$$S \in u_{i,j}(w) \wedge R_i \circ S_{i,j}(w) \subseteq S \Rightarrow R_i(w) \subseteq S; \tag{4}$$

Thus a frame is called a **BIT**-frame if the above two conditions hold. One of the main results of Liau's paper is the completeness of the logic **BIT** with respect to the given semantics.

Fact 6. *[9] The logic **BIT** is sound and complete w.r.t. the class of all **BIT** frames.*

Although the semantics is appealing and the idea to use neighborhood structures is useful for interpreting the trust modality, nevertheless there are some limitations. In our view the semantics is in some sense too strong and in other respects too weak. The main issue concerns how trust relates to belief. As we have just seen it is an axiom of Liau's system that if an agent i trusts an agent j with respect to a proposition φ and moreover believes that the information that φ has been received from j, then i

Figure 2:

believes φ. Nevertheless under these same conditions it does not follow that i believes that j believes φ, nor even that i believes that j disbelieves $\neg\varphi$. This seems rather counter-intuitive.

To see this, let us consider the model from Figure 2. Let $W = \{s, t\} \cup U$ where U is an arbitrary set. Let $R \subseteq W \times W$ be defined as follows $R = U \times U \cup \{(s, t), (t, t)\}$. Let $R_i = S_{i.j} = S_{j,i} = R$ i.e. all relations are the same as R. Let $u_{i,j}(w) = \{U\}$ for each $w \in W$. Let the proposition p be true at U i.e. $V(p) = U$.

It is easy to check that R is euclidean, transitive and serial, hence each R_i and $S_{i,j}$ is defined correctly. It is left to check that the frame defined in the example satisfies the above mentioned two frame properties, (3) and (4). It is immediate that $u_{i,j}(w) = \bigcap_{v \in R_i(w)} u_{i,j}(v)$ since each point in W has a successor (by each R_i) and $u_{i,j}(w)$ is constantly equal to U for each $w \in W$. Now let us check the other property. We consider two cases:

Case 1 ($w = s$ or $w = t$) In this case $R_i \circ S_{i,j}(w) = \{t\}$ while $u_{i,j}(w) = U$ and $\{t\} \nsubseteq U$ hence the implication (4) is satisfied.

Case 2 ($w \in U$) In this case $R_i \circ S_{i,j}(w) = U$ and also $u_{i,j}(w) = U$ which means that the left hand side of the implication (4) is satisfied. But in this case the right hand side is also trivially satisfied as $R_i(w) = U$. Thus we have checked that the frame satisfies all required properties. Therefore the model $\mathcal{M} = (W, R_i, S_{i,j}, u_{i,j}, V)$ belongs to the class of the **BIT**-models. Below in Proposition 7 we present some properties of the model $\mathcal{M}$. We will make use of an additional operator introduced by Liau, called *cautious* trust and denoted by $T^c_{i,j}p$. It is defined by him as follows.

$$T^c_{i,j}p := \Box_i((I_{i,j}p \to \Box_j p) \wedge (\Box_j p \to p)) \tag{5}$$

Proposition 7.
- *$\mathcal{M}, s \nVdash T_{i,j}p \to \Box_i \Diamond_j p$;*
- *$\mathcal{M}, s \nVdash T_{i,j}p \to \Box_i \Box_j p$;*
- *$\mathcal{M}, s \nVdash (T_{i,j}p \to T^c_{i,j}p) \to \Box_i \Diamond_j p$;*
- *$\mathcal{M}, s \nVdash (T_{i,j}p \to T^c_{i,j}p) \to \Box_i \Box_j p$;*

Proof. That $\mathcal{M}, s \Vdash T_{i,j}p$ is clear inasmuch as $V(p) = u_{i,j}(s) = U$. So for the first two items it is left to show that the right hand sides of the implications are falsified. But this is trivial since the only R_i-successor of s is t and besides t is the only R_j - successor of t and $t \Vdash \neg p$. Now for the second two items we are done if we show that $s \Vdash T_{i,j}p \to T^c_{i,j}p$. This amounts to showing that $s \Vdash T^c_{i,j}p$ since it holds that $s \Vdash T_{i,j}p$. This by definition of $T^c_{i,j}$ means that every R_i-successor of s should satisfy $(I_{i,j}p \to \Box_j p) \wedge (\Box_j p \to p)$. We know that t is the only successor of s therefore we should check that $t \Vdash (I_{i,j}p \to \Box_j p) \wedge (\Box_j p \to p)$. Now as $t \nVdash p$ and as t is the only R - successor of w (and $R = R_j = S_{i,j}$) we have that $t \nVdash I_{i,j}p$ and $t \nVdash \Box_j p$. Hence each implication in the conjunction is satisfied. $\Box$

The following corollary directly follows from Fact 6 and Proposition 7.

Corollary 8. *The formulas* $T_{i,j}p \to \Box_i \Diamond_j p$, $T_{i,j}p \to \Box_i \Box_j p$, $(T_{i,j}p \to T^c_{i,j}p) \to \Box_i \Diamond_j p$ *and* $(T_{i,j}p \to T^c_{i,j}p) \to \Box_i \Box_j p$ *are not provable in the logic* ***BIT****.*

We note that the antecedent of the third and fourth implications, namely $(T_{i,j}p \to T^c_{i,j}p)$, is considered by Liau as an additional axiom (C5 in [9]) as a means to capture the cautious trust notion. So even cautious trust in his system does not imply an assumption of honesty on the part of the trusted agent. On the other hand, as Liau points out, it does imply the credibility condition $T^c_{i,j}p \to \Box_i p$ assuming that $\Box_i I_{ij} p$.

In the remainder of the paper we study a system in which this behaviour of trust and cautious trust from **BIT** is no longer to be found. In particular we introduce axioms which in an intuitive way connect trust with the beliefs of agents in such a way that forms of honesty are assumed while credibility is not. In the next section we define two logics **L1** and **L2** of trust where the information acquisition operator no longer plays a role and we only focus on the interconnection of belief and trust. Despite this, due to the new axioms, we have a strong connection between trust and belief both on syntactic and semantic levels.

4 Logics for Trust

Our language will be an extension of **KS** for the multi-agent case together with added modalities for trust. We take from [9] the idea of using neighbourhood structures, but we use a simpler language without an information operator I_{ij} and therefore concentrate only on the interrelation between trust and belief. Based on this language we study two different logics, giving a semantics for each and proving completeness theorems. As mentioned, we will assume that that the doxastic properties of agents follow the **KS** axioms and not those of **KD45**.

4.1 Syntax and Semantics for the Logic of Trust L1

The normal modal logic **L1** for a set of agents A is defined in a signature with an infinite set of propositional letters $Prop = \{p, q, r, ..\}$ and connectives $\wedge, \neg, \Box_i, T_{i,j}$ where i, j belong to the set A. Formulas are built up inductively according to the grammar:

$$\phi ::= p \mid \neg\phi \mid \phi \wedge \phi \mid \Box_i \phi \mid T_{i,j}\phi;$$

where p ranges over propositional symbols in $Prop$ and i, j belong to the set of agents A. The logical symbols '$\top$' and '$\bot$', and the additional connectives such as '$\vee$', '$\to$' and '$\leftrightarrow$' and the dual modality '$\Diamond$', are defined in a standard way.

Axioms for the logic **L1** are given by:

- all classical tautologies;
- each $\Box_i$ satisfies axioms of the logic **KS**;
- $\vdash T_{i,j}p \leftrightarrow \Box_i T_{i,j}p$;
- $\vdash T_{i,j}p \to \Box_i \Diamond_j p$;

for every distinct pair of agents i, j from A.

The rules of inference are: modus ponens, substitution, necessitation for each modality $\Box_i$ where $i \in A$ and the following rule

$$\frac{\vdash p \leftrightarrow q}{\vdash T_{i,j}p \leftrightarrow T_{i,j}q} \tag{6}$$

for each $T_{i,j}$ with $i, j \in A$.

The intended interpretation of $T_{i,j}p$ carries the following idea: agent i trusts agent j about the claim p. In this setting the last two axioms seem very intuitive, in particular: agent i trusts agent j about the claim p if and only if agent i believes that he trusts agent j about p. Hence trust does not contradict one's belief in trust, and if agent i trusts (the honesty of) agent j about p then i believes that j does not disbelieve p. These are the only restrictions we have on the interrelation between trust and belief and appear to be very natural.

A semantics for $\mathbf{L1}$ is provided by frames with weakly transitive and symmetric relations for interpreting belief modalities and neighborhood functions for interpreting trust modalities.

Definition 9. *An $\mathbf{L1}$-frame $\mathfrak{F}$ is a tuple $(W, R_i, u_{i,j})$ for $i, j \in A$ and $i \neq j$; where:*

W is a set of possible worlds;

$R_i \subseteq W \times W$ are weakly transitive and symmetric relations for each $i \in A$;

$u_{i,j} : W \to \mathcal{PP}(W)$ are functions (neighborhood maps), such that the following conditions are satisfied:

$u_{i,j}(w) = \bigcap_{v \in R_i(w)} u_{i,j}(v)$, for every $i, j \in A$ where $i \neq j$; and

$(\forall w, u, U)((wR_iu$ and $U \in u_{i,j}(w)) \Rightarrow (\exists v$ such that uR_jv and $v \in U))$;

An $\mathbf{L1}$-model is a pair $\mathcal{M} = (\mathfrak{F}, V)$, where $\mathfrak{F}$ is a $\mathbf{L1}$-frame and $V : Prop \to \mathcal{P}(W)$ is a valuation function.

To get a better feeling for the last two conditions in the definition of $\mathbf{L1}$-models let us consider an example. For simplicity we describe just a fragment of a model. Let $W = U_1 \cup U_2 \cup U_3 \cup U_4 \cup \{w, v, u\}$ and let $R_i = \{(w, v), (w, u)\}$. We don't specify in more detail the elements of each set U_i since they are not relevant for understanding the conditions. Let $u_{i,j}(w) = \{U_1, U_2\}$, $u_{i,j}(v) = \{U_1, U_2, U_3\}$ and $u_{i,j}(u) = \{U_1, U_2, U_4\}$. Then the condition $u_{i,j}(w) = \bigcap_{v \in R_i(w)} u_{i,j}(v)$ is simplified to the following

$$u_{i,j}(w) = u_{i,j}(v) \cap u_{i,j}(z).$$

This holds for our example since $\{U_1, U_2\} = \{U_1, U_2, U_3\} \cap \{U_1, U_2, U_4\}$.

To visualize the other condition let us take another model where $W = U \cup \{u, w\}$, $R_1 = \{(w, u)\}$ and $R_2 = \{(u, v)\}$, where v is an element of U. Then the condition $(\forall w, u, U)((wR_iu$ and $U \in u_{i,j}(w)) \Rightarrow (\exists v$ such that uR_jv and $v \in U))$ is satisfied for the specific w and u from our model.

Definition 10. *Satisfaction of a formula in a given $\mathbf{L1}$-model $\mathcal{M} = (\mathfrak{F}, V)$ and a point $w \in W$ is defined inductively as follows:*
$w \Vdash p$ *iff* $w \in V(p)$,
$w \Vdash \neg\alpha$ *iff* $w \nVdash \alpha$,
$w \Vdash \alpha \wedge \beta$ *iff* $w \Vdash \alpha$ *and* $w \Vdash \beta$,
$w \Vdash \Box_i\alpha$ *iff* $(\forall w' \in W)(wR_iw' \Rightarrow w' \Vdash \alpha)$,
$w \Vdash T_{i,j}\alpha$ *iff* $|\alpha| \in u_{i,j}(w)$. *Here $|\alpha|$ denotes the truth set of α i.e.* $\{v \in W \mid v \Vdash \alpha\}$.

A formula is valid in a given $\mathbf{L1}$-model if it is true at every point of the model. A formula is valid in an $\mathbf{L1}$-frame if it is valid in every model based on the frame. A formula is valid in a class of $\mathbf{L1}$-frames if it is valid in every frame in the class.

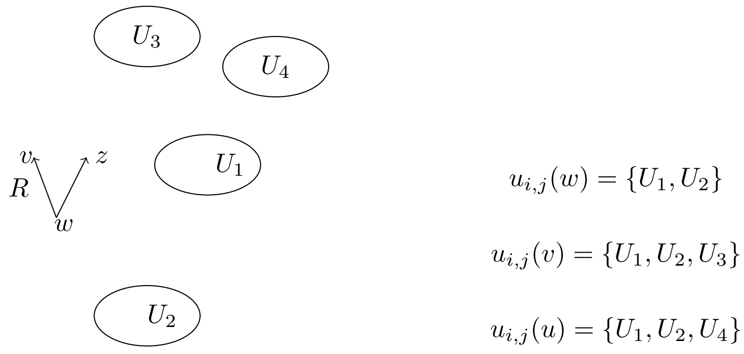

Figure 3:

Figure 4:

Theorem 11. *The logic* $\mathbf{L1}$ *is sound and complete with respect to the class of all* $\mathbf{L1}$*-frames.*

Proof. The soundness easily follows by direct check. For completeness, the proof is standard and therefore we just give a sketch.

Let W be the set of all maximally consistent subsets of formulas in a logic $\mathbf{L1}$. Let us define the relations R_1 and R_2 on W in the following way: For every $\Gamma, \Gamma' \in W$ we define $\Gamma R_i \Gamma'$ iff $(\forall\alpha)(\Box_i\alpha \in \Gamma \Rightarrow \alpha \in \Gamma')$, where $i \in A$. The following lemma is proved in [5] when proving completeness of the logic $\mathbf{KS}$. It also directly follows from the Sahlqvist theorem and the observation that axioms of $\mathbf{KS}$ characterize the class of all weakly-transitive and symmetric frames.

Lemma 12. *[5] Each* R_i *is weakly-transitive and symmetric.*

So far we defined a set W with weakly transitive and symmetric relations R_i on it. Now we define functions $u_{i,j}$ in the following way:

$$u_{i,j}(\Gamma) = \{\{\Gamma' | \phi \in \Gamma'\} \mid T_{i,j}\phi \in \Gamma\}.$$

It immediately follows that $u_{i,j}$ are functions defined from W to $\mathcal{PP}(W)$. The valuation V is defined in the following way: $V(\Gamma) = \{p \in Prop \mid p \in \Gamma\}$.

Lemma 13 (*Truth*)**.** *For every formula* $\alpha \in \mathbf{L1}$ *and every point* $\Gamma \in W$ *of the canonical model, the following equivalence holds:* $\Gamma \Vdash \alpha$ *iff* $\alpha \in \Gamma$.

Proof. The proof goes by induction on the length of formula. Base case follows immediately from the definition of valuation. Assume for all $\alpha \in \mathbf{L1}$ with length less then k holds: $\Gamma \Vdash \alpha$ iff $\alpha \in \Gamma$.

Let us prove the claim for α with length equal to k. If α is conjunction or negation of two formulas then the result easily follows from the definition of satisfaction relation and the properties of maximal consistent sets, so we skip the proofs. Assume $\alpha = \Box_i\beta$ and assume $\Gamma \Vdash \alpha$. Take a set $B = \{\gamma \mid \Box_i\gamma \in$

$\Gamma\} \cup \{\neg\beta\}$. The sub-claim is that B is inconsistent. Assume not, then there exists $\Gamma' \in W$ such that $\Gamma' \supseteq B$. This by definition of the relation R_i means that $\Gamma R_i \Gamma'$. Now as $\neg\beta \in \Gamma'$, by the induction hypothesis we get $\Gamma' \Vdash \neg\beta$. Hence we obtain a contradiction with our assumption that $\Gamma \Vdash \Box_i\beta$. So B is inconsistent. This means that there exist $\gamma_1, \gamma_2, ...\gamma_n \in B$ such that $\vdash \gamma_1 \wedge \gamma_2 \wedge .. \wedge \gamma_n \to \beta$. Applying the necessitation rule for $\Box_i$ we get $\vdash \Box_i\gamma_1 \wedge .. \wedge \Box_i\gamma_n \to \Box_i\beta$ so $\Gamma \vdash \Box_i\beta$, hence it follows that $\Box_i\beta \in \Gamma$.

We just showed the left-to-right direction of our claim for $\alpha = \Box_i\beta$. For the right-to-left implication assume that $\Box_i\beta \in \Gamma$. From the definition of R_i for every Γ' with $\Gamma R_i \Gamma'$ we have $\beta \in \Gamma'$. From this by the inductiion hypothesis it follows that $\Gamma' \vdash \beta$. So we conclude that $\Gamma \Vdash \Box_i\beta$.

Now assume $\alpha = T_{i,j}\phi$. Assume $\Gamma \Vdash T_{i,j}\phi$. By definition this means that $|\phi| \in u_{i,j}(\Gamma)$. Hence there exists β such that $\{\Gamma''|\beta \in \Gamma''\} = |\phi|$ with $T_{i,j}\beta \in \Gamma$. This means that we have $\vdash \beta \leftrightarrow \phi$ in $\mathbf{L1}$. Hence by the rule for trust modality we have $\vdash T_{i,j}\beta \leftrightarrow T_{i,j}\phi$. But the last implies that $T_{i,j}\phi \in \Gamma$.

Conversely assume that $T_{i,j}\phi \in \Gamma$ this implies that $\{\Gamma''|\phi \in \Gamma''\} \in u_{i,j}(\Gamma)$. Now by the induction hypothesis we know that $\Gamma'' \Vdash \phi$ iff $\phi \in \Gamma''$ hence $|\phi| \in u_{i,j}(\Gamma)$. Hence $\Gamma \Vdash T_{i,j}\phi$. □

Now let us show that the model we constructed falls into the class of $\mathbf{L1}$-models. Let us first show the equality:

$$u_{i,j}(\Gamma) = \bigcap_{\Gamma' \in R_i(\Gamma)} u_{i,j}(\Gamma').$$

Assume $X \in u_{i,j}(\Gamma)$. This means that X is of the form $\{\Gamma''|\phi \in \Gamma''\}$ for some ϕ with $T_{i,j}\phi \in \Gamma$. Because of the $\mathbf{L1}$ axioms and because Γ is a maximally consistent set, it follows that $\Box_i T_{i,j}\phi \in \Gamma$. From this we can conclude that $T_{i,j}\phi \in \Gamma'$ for every $\Gamma' \in R_i(\Gamma)$. Now by the definition of $u_{i,j}$ this means that $\{\Gamma''|\phi \in \Gamma''\} \in u_{i,j}(\Gamma')$ and as Γ' was arbitrary member of $R_i(\Gamma)$, we get that $X = \{\Gamma''|\phi \in \Gamma''\} \in \bigcap_{\Gamma' \in R_i(\Gamma)} u_{i,j}(\Gamma')$.

Conversely assume some set $X \subseteq W$ belongs to $\bigcap_{\Gamma' \in R_i(\Gamma)} u_{i,j}(\Gamma')$. By definition this means that there exists a formula ϕ such that $T_{i,j}\phi \in \bigcap_{\Gamma' \in R_i(\Gamma)} \Gamma'$ and $X = \{\Gamma''|\phi \in \Gamma''\}$. Now, since $(\forall\Gamma')(\Gamma R_i \Gamma' \Rightarrow T_{i,j}\phi \in \Gamma')$, by the truth lemma we obtain that $(\forall\Gamma')(\Gamma R_i \Gamma' \Rightarrow \Gamma' \Vdash T_{i,j}\phi)$. Hence $\Gamma \Vdash \Box_i T_{i,j}\phi$. Applying the axioms for the trust modality we can infer that $\Gamma \Vdash T_{i,j}\phi$ and hence $X \in u_{i,j}(\Gamma)$. This completes the proof.

It remains to show the last property, ie., that for every $\Gamma, \Gamma' \in W$ and for every $U \subseteq W$ from $\Gamma R_i \Gamma'$ and $U \in u_{i,j}(\Gamma)$, there exists Γ'' such that $\Gamma' R_j \Gamma''$ and $\Gamma'' \in U$. Assume $\Gamma R_i \Gamma'$ and $U \in u_{i,j}(\Gamma)$. The last, by the definition of $u_{i,j}$, implies that there is a formula ϕ such that $T_{i,j}\phi \in \Gamma$ and $U = \{\Sigma \in W \mid \phi \in \Sigma\}$. By the axiom $T_{i,j}\phi \to \Box_i\Diamond_j\phi$, we get $\Box_i\Diamond_j\phi \in \Gamma$ and, since $\Gamma R_i \Gamma'$, we can infer that $\Diamond_j\phi \in \Gamma'$. Now by a standard argument the set $\{\psi \mid \Box_j\psi \in \Gamma'\} \cup \{\phi\}$ is consistent. Therefore there is a maximal consistent set Γ'' containing it. Thus we infer that $\Gamma' R_j \Gamma''$ and $\phi \in \Gamma''$, which; implies that $\Gamma'' \in U$.

□

4.2 Syntax and Semantics for the Logic of Trust L2

Let us consider another logic of trust in which we vary the assumption of honesty. The only difference with the logic of trust from previous section is in the last axiom $T_{i,j}p \to \Box_i\Diamond_j p$. We replace this axiom by the axiom $T_{i,j}p \to \Box_i\Box_j p$ which informally says that if agent i trusts agent j about p then i believes that j believes that p. The language is the same as in previous section. The axioms for the logic $\mathbf{L2}$ are thus given by:

- all classical tautologies;
- each $\Box_i$ satisfies axioms of the logic $\mathbf{KS}$;

- $\vdash T_{i,j}p \leftrightarrow \Box_i T_{i,j}p$;
- $\vdash T_{i,j}p \rightarrow \Box_i \Box_j p$;

for every distinct pair of agents i, j from A;

As before the rules of inference are: modus ponens, substitution, necessitation for each modality $\Box_i$ where $i \in A$ and the rule (6):

$$\frac{\vdash p \leftrightarrow q}{\vdash T_{i,j}p \leftrightarrow T_{i,j}q}$$

for each $T_{i,j}$ with $i, j \in A$.

Semantics is provided again by frames with weakly transitive and symmetric relations for interpreting belief modalities and neighbourhood functions for interpreting trust modalities.

Definition 14. *A* $\mathbf{L2}$*-frame* $\mathfrak{F}$ *is a tuple* $(W, R_i, u_{i,j})$ *for* $i \in A$ *and* $i \neq j$*, where:*
W *is a set of possible worlds;*

$R_i \subseteq W \times W$ *are weakly transitive and symmetric relations for each* $i \in A$*;*

$u_{i,j} : W \rightarrow \mathcal{PP}(W)$ *are functions (neighbourhood maps), such that the following conditions are satisfied:*

$u_{i,j}(w) = \bigcap_{v \in R_i(w)} u_{i,j}(v)$*, for every* $i, j \in A$ *where* $i \neq j$*; and*

$(\forall U \subseteq W, \forall w \in W)(U \in u_{i,j}(w) \rightarrow R_i \circ R_j(w) \subseteq U)$ *where* $R_i \circ R_j(w)$ *is a set of points reachable from* w *by the relation* $R_i \circ R_j$*;*

An $\mathbf{L2}$*-model is a pair* $\mathcal{M} = (\mathfrak{F}, V)$*, where* $\mathfrak{F}$ *is an* $\mathbf{L2}$*-frame and* $V : Prop \rightarrow \mathcal{P}(W)$ *is a valuation function. Satisfaction and validity of a formula in a given* $\mathbf{L2}$*-model* $\mathcal{M} = (\mathfrak{F}, V)$ *and at a point* $w \in W$ *is defined in a standard way.*

The first condition on neighbourhood maps is the same as for the logic $\mathbf{L1}$. To make the second condition more intuitive let us again consider an example, see the diagram below. Let $W = U \cup \{u, w, v, t\}$, $R_1 = \{(w, u)\}$ and $R_2 = \{(u, v), (u, t)\}$ where v is an element of U while t is not, let $u_{1,2}(w) = U$. Then the condition $(\forall U \subseteq W, \forall w \in W)(U \in u_{i,j}(w) \rightarrow R_i \circ R_j(w) \subseteq U)$ is not satisfied because $R_1 \circ R_2(w) = \{v, t\}$ and $\{v, t\} \nsubseteq U$; while it would be satisfied if t were an element of U.

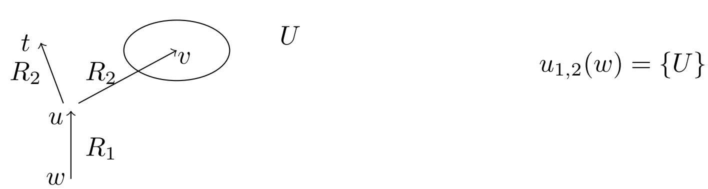

Figure 5:

Theorem 15. *The logic* $\mathbf{L2}$ *is sound and complete with respect to the class of all* $\mathbf{L2}$*-frames.*

Proof. The proof is analogous to that for the logic $\mathbf{L1}$ therefore we just present the case involving the new axiom $T_{i,j}p \rightarrow \Box_i \Box_j p$, i.e., we show that the canonical model for $\mathbf{L2}$ has the property that for every $U \subseteq W$ and for each point $\Gamma \in W$ if $U \in u_{i,j}(\Gamma)$ then $R_i \circ R_j(\Gamma) \subseteq U$. Assume that $U \in u_{i,j}(\Gamma)$. This by definition means that there is a formula ϕ such that $T_{i,j}\phi \in \Gamma$ and $U = \{\Sigma \in W \mid \phi \in \Sigma\}$. By axiom $T_{i,j}\phi \rightarrow \Box_i \Box_j \phi$ we get $\Box_i \Box_j \phi \in \Gamma$. This implies that for each R_i - successor Γ' of Γ we have $\Box_j \phi \in \Gamma'$ which on its own implies that for each R_j-successor Γ'' of Γ' we have $\phi \in \Gamma''$. Hence for each $\Gamma'' \in R_i \circ R_j(\Gamma)$ it holds that $\phi \in \Gamma''$. The last by the definition of $u_{i,j}$ means that $\Gamma'' \in U$.

$\Box$

5 Conclusions

We have made a contribution to the study of the logic of trust by considering some systems involving trust and belief and their interrelations. We have seen that in an interesting approach adopted by Liau in [9] one agent's trust in another does not generally imply an assumption of honesty, though it does, given a suitable information flow, imply credibility. In our alternative logics for trust, we have considered two types of assumptions of honesty that are independent of any credibility assumption. In one case the assumption is that if i trusts j about a proposition p, then i believes that j is not deceitful and does not disbelieve p. In the other case, there is a positive assumption of honesty or sincerity in the sense that i believes that j believes p. Using the logic $\mathbf{KS}$ to model beliefs, these two assumptions are actually independent. In a stronger doxastic logic containing axiom $\mathbf{D}$, the former assumption would be a consequence of the latter. Although we have preferred to take $\mathbf{KS}$ as a base logic here, similar completeness results would hold if $\mathbf{KSD}$ or even $\mathbf{KD45}$ would be used.

It remains for future work to be seen how these logics might accommodate other types of trust/belief relations, such as credibility and other conditions considered for example by [3].

References

[1] Baltag A., Renne B. and Smets S., The Logic of Justified Belief, Explicit Knowledge and Conclusive Evidence. *Annals of Pure and Applied Logic* 165(1): 49-81 (2014).

[2] Blackburn P., van Benthem J., Wolter, F., *Handbook of Modal Logic*, Elsevier, 2002.

[3] Demolombe, R., To trust information sources: a proposal for a modal logical framework, in, C. Castelfranchi, Y-H. Tan, (eds), *Trust and Deception in Virtual Societies*, Kluwer, 2001.

[4] Dundua, B. and Uridia, L., Trust and Belief Interrelation-Modal Logic Approach. In: Proceedings of the 10th International Workshop on Agreement Technologies, Bahia Blanca, Argentina, 2010.

[5] Esakia, L., Weak transitivity-restitution. *Logical Studies*, vol 8, 244-255 (2001).

[6] Goranko, V., Modal definability in enriched language, *Notre Dame Journal of Formal Logic*, 31, 81-105, (1990).

[7] Halmos, P. R., *Algebraic Logic*. Chelsea Publishing Company, New York, 1962.

[8] Herzig, A., Lorini E., Hubner J. F., and Vercouter L., A logic of trust and reputation. *Logic Journal of IGPL*, 18(1), 214-244, (2009).

[9] Liau., Belief, information acquisition, and trust in multi-agent systems - A modal logic formulation, *Journal of Artificial Intelligence*, 149, 31- 60, (2003).

[10] Pearce, D. and Uridia, L., An Approach to Minimal Belief via Objective Belief, in *IJCAI 2011, Proc 22nd Int Joint Conf on Artificial Intelligence, Barcelona, 2011*, pp.1045–1050, 2011.

[11] Pearce, D and Uridia, L, The Topology of Common Belief, in A. Herzig, E. Lorini (eds.), *The Cognitive Foundations of Group Attitudes and Social Interaction*, Studies in the Philosophy of Sociality 5, DOI 10.1007/978-3-319-21732-1 7, Springer, to appear.

[12] Ramchurn, S.D., Huynh, T.D. and Jennings, N.R., Trust in multi-agent systems, *The Knowledge Engineering Review*, 19:1, 25, 2004.

[13] de Rijke, M., The Modal Logic of Inequality, *J. Symbolic Logic*, 57, 566-587, (1992).

[14] Rodenhauser, B., A Matter of Trust. PhD thesis, University of Amsterdam, 2014

[15] Sakama, C, Caminada, M and Herzig, A, A Formal Account of Dishonesty. *Logic Journal of the IGPL* 23(2): 259–294 (2015). http://dare.uva.nl/document/2/141925.

[16] Segerberg, K., ”Somewhere else” and ”some other time”, In: Wright and wrong: mini-essay in honor of G.H.Wright, Publ. the group in logic and methodology of Real Finland, 61-64, 1976.

[17] Smullyan, R.M., *What Is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles*, Dover, 2011.