A minimal logic of trust

Andreas Herzig
CNRS, IRIT, Toulouse

Abstract

We express Castelfranchi and Falcone's semi-formal definition of trust in a logic of belief and agency. In order to identify the key mechanisms and to keep complexity low we keep the logic as simple as possible.

1 Introduction

Any society can be understood as a web of truster-trustee relations. Consider for example a conference banquet: the conference participants trust the organiser to ensure a decent dinner; the organiser trusts the caterer to deliver a decent conference banquet; the organiser trusts the participants to attend the dinner; the organiser trusts the administration of her lab to take car of the payment of the caterer; the caterer trusts the university to pay the bill; etc. Our example involves both humans and institutions, and almost everybody is both a truster and a trustee.

The concept of trust is studied since long in psychology, sociology, and economics. Recently the opaqueness of deep learning-based systems has highlighted the importance of trusted AI systems and has triggered substantial research efforts in particular in the EU. According to the EU AI Act,

> "In a context of rapid technological change, we believe it is essential that trust remains the bedrock of societies, communities, economies and sustainable development. We therefore identify Trustworthy AI as our foundational ambition, since human beings and communities will only be able to have confidence in the technology's development and its applications when a clear and comprehensive framework for achieving its trustworthiness is in place."

The aim of trustworthy AI is to come up with structures such as trust metrics that institutionalise trust. Interestingly, it seems that it is taken for granted that trust in an AI system is a consequence of its trustworthiness. However, as already pointed out by Cristiano Castelfranchi this fails to hold: a trustworthy AI system might not be trusted by the intended human user. For example, during the COVID-19 pandemic many contact tracing apps were judged by experts to be trustworthy w.r.t. the standards of privacy, but many citizens nevertheless failed to trust these applications. Another questionable point is that the EU guidelines on trustworthy AI view trust and trustworthiness as something that is good 'by nature' and should be promoted. This seems to neglect that, by its very nature, trust also comes with dependence, uncertainty and risk and that distrust is a better attitude in many cases: one may consider that the attitudes of mistrust and distrust are appropriate attitudes in many circumstances. This indicates that it would be good to complement the EU strategy by structures institutionalising mistrust and distrust (Reinhardt, 2023).

We here account for the relation of trust of one individual i (the truster) in another individual j (the trustee). The trustee is not always a particular human being: other kinds of trustees are institutions such as states, banks, and companies; in trustworthy AI the trustees are algorithms (and their designers and the platforms using them). We here understand that the trustee is either a human or an artificial agent. We provide a formal logical analysis of a well-known theory of trust: that of Castelfranchi and Falcone (C&F henceforth). Such formal systems have been proposed before, e.g. by Jones and Firozabadi (2001); Jones (2002); Demolombe (2004); Lorini and Demolombe (2008); Herzig et al. (2010); Demolombe (2017). They are however highly complex and only give little formal results about the logic beyond completeness of the axiomatisation. In contrast, in this paper we design a minimalist logic that provides what we claim to be the simplest logic of trust. This allows us to give several formal results about our logic, in particular complexity results. We believe such formal models and results to be relevant in order to clarify concepts and compare

existing computational models which are mainly quantitative, have varying focus (trust, reputation, reliability, sincerity, service quality,...), and are difficult to compare. Beyond, it can be expected to contribute to the verifiability of implemented trust systems and to artificial agents with built-in capabilities to reason about trust.

2 The logical form of the trust relation

Many approaches in AI and MAS consider that the trust relation has only two arguments: a truster i and a trustee j. This is however not enough if trust is about several topics. For example, if I trust my physician to to treat my flu then I do not necessarily trust her to repair my car. Hence the trust relation should have an *action* of the trustee as an argument. We note that actions have to be understood in a large sense because they should include refraining from acting. For example, I trust the French smartphone app `TousAntiCovid` to store my $\geq$15mn contacts, to check if it is known that there was a Covid contact, to warn that contact if I am infected and ask the application to do so; and I also trust *TousAntiCovid* to *refrain* from performing actions such as warning contacts without asking me and revealing my position to the police. Finally, I *do not* trust *TousAntiCovid* to provide advice about vaccines or to warn me about flu contacts.

Is trust a 3-ary relation $\mathit{Trust}(i, j, \mathit{act})$ between a truster i, a trustee j, and some action act of j? That relation was called "core trust" in the literature (Jones, 2002; Castelfranchi and Falcone, 1998). However, C&F have argued that core trust is not enough (Falcone and Castelfranchi, 2001) and that trust is only about actions that are relevant for the trustee: there is something at stake for i, and j's action should promote i's goals or avoid harm to i. Hence trust implies risk: i is vulnerable and depends on j. The presence of a goal argument also leads us to distinguish trust that j *achieves* something desirable for i from trust that j *maintains* something desirable for i.

These considerations lead to the analysis of trust as a 4-ary relation $\mathit{Trust}(i, j, \mathit{act}, \varphi)$ where act is an action of i and φ is a goal of j. It reads "i trusts j to perform act in order to achieve φ". A semi-formal analysis identifying the main ingredients of that relation was provided by C&F in several publications that became standard references in the AI and multi-agent systems literature (Castelfranchi and Falcone, 1998; Falcone and Castelfranchi, 2001). In C&F theory trusting is believing without knowing: trust is a belief of the truster i that may turn out to be wrong when the trustee j does not act as expected. These beliefs come with a complex evaluation of the trustee j concerning j's *ability* to perform an action ("j's competence"), j's *willingness* to perform an action ("j's disposition"), and *dependence*, that is, j's action will fullfil a goal of the truster. More precisely, C&F's define $\mathit{Trust}(i, j, \mathit{act}, \varphi)$ as the conjunction of the following:

- i has the goal that φ be true;
- i believes that j's successful performance of act fulfills φ;
- i believes that j is able to perform act;
- i believes that j is willing to perform act.

The concept of trust is thus reduced to 6 more primitive concepts: goal, belief, action, dependence, ability, and willingness. At the heart of the four latter is the concept of *agency*, that should therefore be central in the analysis. It is traditionally viewed as a relation between an individual and a proposition: "individual i is agentive for proposition φ", reformulated by Belnap to "i sees to it that φ" and abbreviated $[i\,\mathrm{stit}]\varphi$ (Belnap et al., 2001). It is fundamental not only in the analysis of trust, but also of causality, responsibility, influence, deception, manipulation, and social emotions such as regret. In order to account for fullfilment we add actions as a third argument and formalise sentences of the kind "i sees to it that φ by doing act", as previously done e.g. in (Herzig and Lorini, 2010).

It is clear from what we have said up to now that logical accounts of trust should tend to be very complex, both conceptually and computationally. Conceptually, the difficulty is to choose among the existing formalisations of the above concepts and to account for the complex interactions between them. The computational difficulties are inherited from the complexity of the logics of each of the above concepts: the satisfiability problem for the standard modal logic of belief KD45 is PSpace-hard as soon as two or more agents are involved (Halpern and Moses, 1992; Fagin et al., 1995); that for the logic of seeing-to-it-that is NExpTime-hard

(Balbiani et al., 2008). Moreover, while model checking is in general a less complex and therefore interesting altenative to satisfiability checking and theorem proving (Halpern and Vardi, 1991), it is a priori unfeasible here because the Kripke models for logics of belief and for logics of agency are already typically too big to be manageable. Our strategy in this paper is to start out from simple fragment of the logic of agency that was recently proposed in (Herzig et al., 2022) and to build a minimal logic of trust on top of it. This will guarantee better computational properties; in particular, the models will be compact enough to make model checking an option.

3 A basic logic of control and attempt

We now introduce a simplified version of the logic of agency based on attempt and control (LACA) of (Herzig et al., 2022). We call that logic LACA$^-$.

The alphabet of the language of LACA$^-$ has a set of agents *Agt* with typical elements $i, j, \ldots$ and a set of propositional variables *Prop* with typical elements $p, q, \ldots$, plus special variables $\mathbf{c}_i p$ and $\mathbf{t}_i p$ where $i \in Agt$ is an agent and $p \in Prop$ a propositional variable. The former reads "agent i controls p" and the latter "agent i attempts to change p". The set of control variables and the set of attempt variables are therefore:

$$\mathbf{Ctrl} = \{\mathbf{c}_i p \ : \ i \in Agt, p \in Prop\},$$
$$\mathbf{Tmpt} = \{\mathbf{t}_i p \ : \ i \in Agt, p \in Prop\}.$$

The set of *atoms* is $Prop \cup \mathbf{Tmpt} \cup \mathbf{Ctrl}$. We use α, β,... to denote atoms.

The language of LACA$^-$ is built from atoms by means of the standard boolean operators together with two modal operators: a 'next' operator that is borrowed from temporal logic and an operator of historic necessity that is borrowed from the logic of 'seeing-to-it-that'. Formulas are defined by the following grammar:

$$\varphi ::= \alpha \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathtt{X}\varphi \mid \Box\varphi$$

where α ranges over $Prop \cup \mathbf{Tmpt} \cup \mathbf{Ctrl}$. The formula $\mathtt{X}\varphi$ reads "next φ" and $\Box\varphi$ reads "φ is historically necessary". The language of LACA$^-$ is a fragment of the language of stit logic: basically $[i\,\mathrm{stit}]\mathtt{X}p$ can be expressed by $\neg p \wedge \mathbf{c}_i p \wedge \mathbf{t}_i p$, that is, i sees to it that next φ is identified with "p is false, i controls p, and i tries to change p". Similarly, $[i\,\mathrm{stit}]\mathtt{X}\neg p$ is expressed by $p \wedge \mathbf{c}_i p \wedge \mathbf{t}_i p$. Hence agency is about propositional variables p: we cannot express that an agent sees to it that a disjunction is true.[1]

A *valuation* $V \subseteq Prop \cup \mathbf{Tmpt} \cup \mathbf{Ctrl}$ allows us to associate truth values to to atoms, that is, to propositional variables, attempt atoms, and control atoms. We may suppose that control is exclusive (if $\mathbf{c}_i p, \mathbf{c}_j p \in \mathbf{Ctrl}$ then $i = j$) but need not do so. An example of a valuation is $V^{hw} = \{\mathbf{t}_1 h, \mathbf{t}_1 w, \mathbf{c}_1 h, \mathbf{c}_2 w\}$ where w stands for "the window is closed" and h for "the heating is on". In that valuation, agent 1 tries to change w but does not have the ability to do so.

A given valuation determines a *successor*: when $\mathbf{c}_i p$ and $\mathbf{t}_i p$ are true then p changes truth value; succesful attempts are abandoned (cf. achievement goals); and the rest is preserved. The formal definition uses the notion of *changeset* of a valuation V, defined as:

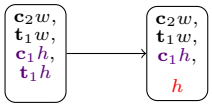
$$V^{\pm} = \{p \in Prop \ : \ \text{there is } i \in Agt \text{ such that } \mathbf{c}_i p, \mathbf{t}_i p \in V\}.$$

The elements of $V^{\pm}$ are the propositional variables whose truth values are going to be flipped. Then the successor of a valuation is define by:

$$\begin{aligned} succ(V) = \ & \{p \in Prop \ : \ p \in V \text{ and } p \notin V^{\pm}\} \cup \\ & \{p \in Prop \ : \ p \notin V \text{ and } p \in V^{\pm}\} \cup \\ & \{\mathbf{c}_i p \in \mathbf{Ctrl} \ : \ \mathbf{c}_i p \in V\} \cup \\ & \{\mathbf{t}_i p \in \mathbf{Tmpt} \ : \ \mathbf{t}_i p \in V \text{ and } p \notin V^{\pm}, i \in Agt\}. \end{aligned}$$

For example, the successor of the above example valuation V^{hw} is depicted below:

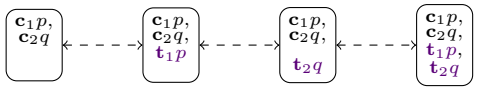
[1] The logic of (Herzig et al., 2022) has further features that we do not consider her for the sake of simplicity, namely the temporal operator "henceforth" of LTL and higher-order attempt and control, that is, control of attempts, attempts to control, etc. For example, if w stands for "the window is closed" then $\mathbf{c}_1\mathbf{t}_2 w$ expresses that 1 can order 2 to open or close the window, and $\mathbf{t}_1\mathbf{t}_2 w$ expresses that 1 tries to make 2 flip her goal about the window.



Historic necessity of φ means that φ is true whatever the agents attempt. This means that we quantify over attempts while keeping facts and abilities constant. We write $V \sim V'$ when V and V' agree on facts and abilities, that is:

$$V \sim V' \quad \text{iff} \quad V \cap \mathit{Prop} = V' \cap \mathit{Prop} \text{ and } V \cap \mathbf{Ctrl} = V' \cap \mathbf{Ctrl}.$$

This is clearly an equivalence relation. It is depicted below for our running example, omitting reflexive and transitive arrows:



The truth conditions for LACA$^-$ formulas are as follows:

$$\begin{array}{lll} V \models \alpha & \text{iff} & \alpha \in V, \text{ for } \alpha \in \mathit{Prop} \cup \mathbf{Tmpt} \cup \mathbf{Ctrl}; \\ V \models \mathtt{X}\varphi & \text{iff} & \mathit{succ}(V) \models \varphi; \\ V \models \Box\varphi & \text{iff} & \text{for every } V' \sim V,\ V' \models \varphi. \end{array}$$

and as usual for the boolean connectives. For example, in the valuation $V^{hw} = \{\mathbf{t}_1h, \mathbf{t}_1w, \mathbf{c}_1h\}$ the formulas $\mathtt{X}(h \wedge \neg w \wedge \mathbf{t}_1w)$, $\neg\Box\mathtt{X}h$, $\neg\Box\mathtt{X}\neg h$, $\neg w$, and $\Box\mathtt{X}\neg w$ are true.

We axiomatise the validities of LACA$^-$ by the following schemas:

$$\begin{aligned} \mathtt{X}(\varphi_1 \wedge \varphi_2) &\leftrightarrow \mathtt{X}\varphi_1 \wedge \mathtt{X}\varphi_2; \\ \mathtt{X}\neg\varphi &\leftrightarrow \neg\mathtt{X}\varphi; \\ \mathtt{X}p &\leftrightarrow \left(p \leftrightarrow \neg \bigvee_i (\mathbf{c}_ip \wedge \mathbf{t}_ip)\right); \\ \mathtt{X}\mathbf{c}_ip &\leftrightarrow \mathbf{c}_ip; \\ \mathtt{X}\mathbf{t}_ip &\leftrightarrow \mathbf{t}_ip \wedge \neg\mathbf{c}_ip; \\ \Box(\varphi_1 \wedge \varphi_2) &\leftrightarrow \Box\varphi_1 \wedge \Box\varphi_2; \\ \Box(\ell \vee \varphi) &\leftrightarrow \begin{cases} \ell \vee \Box\varphi & \text{if } \ell = p, \neg p, \mathbf{c}_ip, \neg\mathbf{c}_ip, \\ \Box\varphi & \text{if } \ell = \mathbf{t}_ip, \neg\mathbf{t}_ip, \not\models \ell \vee \varphi, \\ \top & \text{otherwise;} \end{cases} \end{aligned}$$

together with rules of equivalence for $\mathtt{X}$ and $\Box$. These seven equivalences make up a complete set of reduction axioms: every LACA$^-$ formula can be reduced to a formula of propositional logic that is equivalent to the original formula.

Proposition 1. *Every LACA$^-$ formula is equivalent to a boolean combination of atoms.*

Proposition 2. *A boolean combination of atoms is LACA$^-$ satisfiable if and only if it is satisfiable in propositional logic.*

Actually $\Box$ can be viewed as a propositional quantifier. This indicates that LACA$^-$ satisfiability and model checking are both PSpace complete, which compares favourably to NExpTime-completeness of satisfiability of Belnap's stit-logic.

4 Adding belief

We now add beliefs to the picture: we extend the language of LACA$^-$ by modal operators of belief $\mathbf{B}_i$, one per agent $i \in \mathit{Agt}$. We call the resulting language LACAB. Our language has two kinds of formulas: LACA$^-$

formulas, typically noted φ, which can be thought of as being about the physical world (which we therefore view as comprising the agents' attempts); and LACAB formulas, typically noted ψ, which can be thought of as being about the psychological world. The grammar is:

$$\varphi ::= p \mid \mathbf{c}_i p \mid \mathbf{t}_i p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathtt{X}\varphi \mid \Box\varphi;$$
$$\psi ::= \varphi \mid \neg\psi \mid \psi \wedge \psi \mid \mathbf{B}_i\, \psi;$$

where p ranges over *Prop* and i over *Agt*. The formula $\mathbf{B}_i\, \psi$ reads "i believes that ψ". For example, $\mathbf{B}_2\, (\mathbf{c}_1 w \wedge \neg\mathbf{t}_1 w)$ expresses that agent 2 believes that 1 controls the variable w but is not going to act on it.

Kripke models for LACAB are triples $M = \langle W, \{R_i\}_{i\in Agt}, \mathcal{V}\rangle$ where: W is a non-empty set of possible worlds; every $R_i \subseteq W \times W$ is an accessibility relation that is serial, transitive, and euclidean; and

$$\mathcal{V} : w \in W \mapsto \mathcal{V}(w) \subseteq \mathit{Prop} \cup \mathbf{Ctrl} \cup \mathbf{Tmpt}$$

is a valuation function associating LACA$^-$ valuations to each possible world w of W.

The truth conditions are:

$$\begin{array}{lll} M, w \Vdash \varphi & \text{if} & \mathcal{V}(w) \models \varphi, \text{ for formulas } \varphi \text{ of the language of LACA}^-; \\ M, w \Vdash \mathbf{B}_i\, \psi & \text{if} & M, w' \Vdash \psi \text{ for every } w' \text{ such that } wR_iw'; \end{array}$$

and as usual for the boolean connectives.

We conjecture that the complexity LACAB satisfiability is the same as that of LACA$^-$ satisfiability.

5 Definition of trust

We recall that C&F have proposed to reduce the trust predicate to the conjunction:

$$\mathit{Trust}(i, j, \mathit{act}, \varphi) \stackrel{\mathtt{def}}{=} \mathit{Goal}(i, \varphi) \wedge \mathbf{B}_i\, \mathit{Causes}(j, \mathit{act}, \varphi) \wedge \mathbf{B}_i\, \mathit{Can}(j, \mathit{act}) \wedge \mathbf{B}_i\, \mathit{Will}(j, \mathit{act}).$$

Note that in our formulation φ is in the language of attempt and control, which means that we cannot express epistemic goals. This restriction is due to the two layers of the language of LACAB.

First of all, we are going to show how the term *act* and the predicates $\mathit{Causes}(j, \mathit{act}, \varphi)$, $\mathit{Can}(j, \mathit{act})$, and $\mathit{Will}(j, \mathit{act})$ can be expressed in our logic of agency LACA$^-$. We are then going to express $\mathit{Goal}(i, \varphi)$.

We suppose that actions change the truth values of some variables and leave some others unchanged. We therefore identify an action with a couple $\mathit{act} = \langle \mathit{act}^{\pm}, \mathit{act}^{=}\rangle$, with $\mathit{act}^{\pm}, \mathit{act}^{=} \subseteq \mathit{Prop}$. The set $\mathit{act}^{\pm}$ are the variables whose truth value changes, while $\mathit{act}^{=}$ are the variables that are left unchanged. The second component allows us to account for actions maintaining the value of a variable. A natural requirement is that the sets $\mathit{act}^{\pm}$ and $\mathit{act}^{=}$ are disjoint.

To express the predicate $\mathit{Causes}(j, \mathit{act}, \varphi)$, we resort to the very idea of seeing-to-it-that logics: j causes φ by doing *act* exactly when j's performance of *act* has consequence φ whatever the other agents attempt. We capture quantification over all agents' attempts by means of the operator of historical necessity $\Box$:

$$\mathit{Causes}(j, \mathit{act}, \varphi) \stackrel{\mathtt{def}}{=} \Box((\mathit{Can}(j, \mathit{act}) \wedge \mathit{Will}(j, \mathit{act})) \rightarrow \varphi).$$

As to the ability predicate $\mathit{Can}(j, \mathit{act})$, we suppose that j is able to perform *act* if and only if j controls all the variables of *act*:

$$\mathit{Can}(j, \mathit{act}) \stackrel{\mathtt{def}}{=} \bigwedge_{\alpha\in \mathit{act}^{\pm}\cup \mathit{act}^{=}} \mathbf{c}_j \alpha.$$

As to the willingness predicate $\mathit{Will}(j, \mathit{act})$, we suppose that j is willing to perform *act* if and only if (1) j attempts to change all variables of $\mathit{act}^{\pm}$ and (2) does not attempt to change any variable of $\mathit{act}^{=}$:

$$\mathit{Will}(j, \mathit{act}) \stackrel{\mathtt{def}}{=} \bigwedge_{\alpha\in \mathit{act}^{\pm}} \mathbf{t}_j \alpha \wedge \bigwedge_{\alpha\in \mathit{act}^{=}} \neg\mathbf{t}_j \alpha.$$

We finally express the goal predicate $\mathit{Goal}(i, \varphi)$ by adopting Cohen and Levesque's definition in terms of preference (Cohen and Levesque, 1990). We suppose that i's preferred histories are characterised by a formula

$Pref_i$ in the language of LACA. For example, we may state that i prefers that in the next state the heating is on and the window is open as $Pref_i = \mathtt{X}(\mathsf{HeatingOn} \rightarrow \neg\mathsf{WindowOpen})$. We then define

$$Goal(i, \varphi) \stackrel{\mathtt{def}}{=} \mathbf{B}_i \Box(Pref_i \rightarrow \varphi).$$

Therefore i has the goal that φ if φ holds on all possible histories that are preferred by i.

6 Definition of distrust and mistrust

Iin natural language there is no clear distinction between distrust and mistrust. However, several authors in the literature have assumed fine-grained differences between the two. We here adopt the following distinction. Mistrust is a form of ignorance: the truster has no information about the trustee. This is typically due to the absence of prior interaction and comes with the action tendency of search for information. Distrust is a belief that the trustee is untrustworthy and will act in a way that is bad for the truster. This is typically due to past bad interaction and comes with the action tendency of avoidance of interaction.

This leads to the following definitions in our logic LACAB:

$$\mathsf{Mistrust}(i, j, act, \varphi) \stackrel{\mathtt{def}}{=} Goal(i, \varphi) \ \wedge \neg\mathbf{B}_i \left(Causes(j, act, \varphi) \wedge Can(j, act) \wedge Will(j, act)\right);$$

$$\mathsf{Distrust}(i, j, act, \varphi) \stackrel{\mathtt{def}}{=} Goal(i, \varphi) \ \wedge \mathbf{B}_i \left(Causes(j, act, \varphi) \wedge Can(j, act) \wedge \neg Will(j, act)\right).$$

7 Conclusion

We have succeeded in expressing C&F's trust in a minimal logic of agency, as well as distrust and mistrust. The account can be extended straightforwardly to trusting groups, $Trust(i, J, act, \varphi)$, where J is a set of agents: it suffices to extend the definition of the fullfilment predicate $Causes(j, act, \varphi)$ of Section 5 to:

$$Causes(J, act, \varphi) \stackrel{\mathtt{def}}{=} \bigvee_{j \in J} \Box((Can(j, act) \wedge Will(j, act)) \rightarrow \varphi).$$

The two layers of the language of LACAB make that one cannot express epistemic goals. In future work we plan to overcome this limitation. A further perspective is the formalisation of reputation as the belief of a group that the trustee's action will achieve the group goal, as initially proposed in (Herzig et al., 2010).

References

Balbiani, P., Herzig, A., and Troquard, N. (2008). Alternative axiomatics and complexity of deliberative stit theories. *Journal of Philosophical Logic*, 37(4):387–406.

Belnap, N., Perloff, M., and Xu, M. (2001). *Facing the Future: Agents and Choices in Our Indeterminist World*. Oxford University Press, Oxford.

Castelfranchi, C. and Falcone, R. (1998). Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings of the Third International Conference on Multiagent Systems (ICMAS'98)*, pages 72–79.

Cohen, P. R. and Levesque, H. J. (1990). Intention is choice with commitment. *Journal of Artificial intelligence*, 42(2):213–261.

Demolombe, R. (2004). Reasoning about trust: A formal logical framework. In Jensen, C. D., Poslad, S., and Dimitrakos, T., editors, *Trust Management, Second International Conference, iTrust 2004, Oxford, UK, March 29 - April 1, 2004, Proceedings*, volume 2995 of *Lecture Notes in Computer Science*, pages 291–303. Springer.

Demolombe, R. (2017). Trust and agency in the context of communication. *J. Appl. Non Class. Logics*, 27(1-2):140–151.

Fagin, R., Halpern, J. Y., Moses, Y., and Vardi, M. Y. (1995). *Reasoning about Knowledge*. MIT Press.

Falcone, R. and Castelfranchi, C. (2001). Social trust: A cognitive approach. In Castelfranchi, C. and Tan, Y. H., editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer.

Halpern, J. Y. and Moses, Y. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54(3):319–379.

Halpern, J. Y. and Vardi, M. Y. (1991). Model checking vs. theorem proving: A manifesto. In Allen, J. F., Fikes, R., and Sandewall, E., editors, *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91). Cambridge, MA, USA, April 22-25, 1991*, pages 325–334. Morgan Kaufmann. Also in Artificial and Mathematical Theory of Computation 1991: 151-176.

Herzig, A. and Lorini, E. (2010). A dynamic logic of agency I: stit, capabilities and powers. *Journal of Logic, Language and Information*, 19(1):89–121.

Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244. Special Issue "Normative Multiagent Systems".

Herzig, A., Lorini, E., and Perrotin, E. (2022). A computationally grounded logic of 'seeing-to-it-that'. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2648–2654. ijcai.org.

Jones, A. J. I. (2002). On the concept of trust. *Decision Support Systems*, 33(3):225–232.

Jones, A. J. I. and Firozabadi, B. (2001). On the characterisation of a trusting agent - aspects of a formal approach. In Castelfranchi, C. and Tan, Y. H., editors, *Trust and Deception in Virtual Societies*, pages 157–168. Kluwer.

Lorini, E. and Demolombe, R. (2008). From binary trust to graded trust in information sources: A logical perspective. In Falcone, R., Barber, K. S., Sabater-Mir, J., and Singh, M. P., editors, *Trust in Agent Societies, 11th International Workshop, TRUST 2008, Estoril, Portugal, May 12-13, 2008. Revised Selected and Invited Papers*, volume 5396 of *Lecture Notes in Computer Science*, pages 205–225. Springer.

Reinhardt, K. (2023). Trust and trustworthiness in ai ethics. *AI and Ethics*, 3(3):735–744.