Chapter 1
A simple logic of trust based on propositional assignments

Andreas Herzig, Emiliano Lorini, and Frédéric Moisan

Abstract Cristiano Castelfranchi and Rino Falcone introduced an influential cognitive theory of social trust that is based on the concepts of belief, goal, ability, willingness and opportunity. In this paper we propose a simple logic of belief and action that allows to express these concepts. While our logic of belief is standard, our logic of action has a very simple kind actions: actions setting the truth value of a propositional variable to either true or false. We call such actions propositional assignments and argue that our logic provides a framework that is simple but expressive enough to account for Castelfranchi and Falcone's concept. We prove its completeness and give a decision procedure.

1.1 Introduction

Cristiano Castelfranchi and Rino Falcone introduced an influential cognitive theory of social trust that is based on the concepts of belief, goal, ability, willingness and opportunity [1, 6, 2]. According to that theory, "agent i trusts agent j to perform action δ_j to achieve i's goal φ" is defined as follows:

1. i has the goal that φ;
2. i believes that j is able to perform δ_j;
3. i believes that j is willing to perform δ_j;
4. i believes that j has the opportunity to achieve φ by performing δ_j.

We here consider a generalisation of the definition where an agent may trust a group of agents J to perform a joint action δ_J to achieve φ.

Castelfranchi and Falcone's theory of trust was the starting point of the ForTrust project[1] that was funded by the French *Agence Nationale de la Recherche* from 2007 to 2010. One of the aims of ForTrust was to design a formal logical framework for their theory.

Université de Toulouse, IRIT-CNRS, France

[1] www.irit.fr/ForTrust

In previous work we had defined a logic of time, action, belief and choice where the above concepts of goal, belief, ability, willingness and opportunity can be expressed and within which one can therefore formally reason about trust [7]. That logic combined temporal, dynamic and epistemic logic and lead to a rather complex formalism. While we were able to state a completeness result we were not able to prove its decidability or to characterise its complexity.

The aim of the present paper is to simplify our account and in particular to state a decidability result. We introduce a simple logic in the style of dynamic epistemic logics [5].

In what concerns actions our account of action is based on the concepts of *propositional assignment* and of *propositional control*. Basically, the idea is that the agents' actions consist in setting the truth values of a propositional variable to either true or false. In order to be able to set variable p to true an agent must have that action in his repertoire, and likewise for the action of setting p to false. As we have shown in [8] this allows to reason about propositional control in van der Hoek and Wooldridge's sense [10, 9]. The latter have used the logic of propositional control in order to talk about an agent's capability to achieve some property (whatever the other agents do). We moreover integrate protocols prescribing what action is going to take place next. This allows not only to talk about what agents *can do* but also about what they actually *do* (according to the current protocol). In previous work agents acted one at a time; we here move to a more game-theoretic account where all agents act in parallel.

Our concept of belief is in terms of the standard logic of belief KD45: we identify belief with truth in all the worlds that are possible for the agent. We also integrate the concept of weak belief and the concept of agent preference. We identify weak belief with truth in the most plausible worlds among those worlds that are possible for the agent; and we identify preference with truth in the most preferred worlds among those worlds that are possible for the agent.

We show that in our logic we can express Castelfranchi and Falcone's relevant concepts of goal, belief, ability, willingness and opportunity as follows:

1. Agent i has the goal that φ if and only if i prefers that φ will be true at the next time step.
2. Agent j is able to perform an action δ_j (of setting p to true or false) if and only if the performance of δ_j by j is possible.
3. Agent j is willing to perform action δ_j if and only if j prefers that j performs δ_j next.
4. Agent j has the opportunity to achieve φ by doing δ_j if and only if φ is true after every possible performance of δ_j by j, whatever the other agents do.

Note that the last three components are logically independent: an agent may be willing to perform δ_j without being able to perform it, etc.

1.2 DDL−PA: dynamic doxastic logic of propositional assignments

In this section we define syntax and semantics of the basic dynamic doxastic logic of propositional assignments DDL−PA. We give its language and semantics and establish decidability of DDL−PA validity.

1.2.1 Language

Let $\mathbb{P}$ be a finite set of propositional variables and let $\mathbb{I}$ be a finite set of individuals, alias agents. An *assignment* is of the form $p{\leftarrow}\top$ or $p{\leftarrow}\bot$, where p is a propositional variable. The set of all assignments is

$$\mathsf{ASS} = \{p{\leftarrow}\top \ : \ p \in \mathbb{P}\} \cup \{p{\leftarrow}\bot \ : \ p \in \mathbb{P}\}.$$

The simultaneous performance of assignments by agents is an *event*. We model events as mappings from $\mathbb{I}$ to ASS. Hence the set of all events is $\mathsf{EVT} = \mathsf{ASS}^{\mathbb{I}}$. We denote the elements of EVT by δ, δ', etc. The restriction of a mapping δ to a set of agents $J \subseteq \mathbb{I}$ is noted δ_J. Hence the restriction of δ to the set of all agents $\mathbb{I}$ is nothing but δ itself: $\delta_{\mathbb{I}} = \delta$. It is convenient to write δ_i instead of $\delta_{\{i\}}$, δ_{-J} instead of $\delta_{\mathbb{I}\setminus J}$, δ_{-i} instead of $\delta_{-\{i\}}$. (Hence we can identify $\delta(i)$ with δ_i.) Observe that EVT is finite because both $\mathbb{P}$ and $\mathbb{I}$ are finite.

The language of $\mathsf{DDL{-}PA}$ is defined by the following grammar:

$$\varphi ::= p \mid \top \mid \bot \mid \mathsf{Hpn}(\delta_i) \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathsf{Bel}_i\,\varphi \mid \mathsf{Next}\,\varphi \mid \langle\delta\rangle\varphi$$

where p ranges over $\mathbb{P}$, i ranges over $\mathbb{I}$ and δ ranges over the set of events EVT.

$\mathsf{Hpn}(\delta_i)$ reads "agent i is going to perform δ_i", or "δ_i is going to happen next"; $\mathsf{Bel}_i\,\varphi$ reads "i believes that φ"; $\mathsf{Next}\,\varphi$ reads "φ is going to be true next"; $\langle\delta\rangle\varphi$ reads "δ may occur, and φ is going to be true immediately afterwards". The modal operator $[\delta]$ is the dual of the modal operator $\langle\delta\rangle$ that is defined by:

$$[\delta]\varphi \stackrel{\mathtt{def}}{=} \neg\langle\delta\rangle\neg\varphi$$

Hence $[\delta]\varphi$ can be read "if δ occurs then φ is going to be true immediately afterwards".

We define $\mathbb{I}_\varphi$ to be the set of agents of $\mathbb{I}$ occurring in formula φ, and we define $\mathbb{P}_\varphi$ to be the set of propositional variables of $\mathbb{P}$ occurring in φ.

1.2.2 Semantics

An *epistemic PC model* is made up a set of possible worlds plus mappings associating to every possible world a valuation, a next state function, a repertoire function, and a belief state per agent. Formally, they are quadruples of the form $M = (W, B, R, N, V)$, where:

- W is a nonempty set of possible worlds;
- $B : W \longrightarrow (\mathbb{I} \longrightarrow 2^W)$ associates to every possible world w and to every agent i the set of worlds that are possible for i at w;
- $R : W \longrightarrow ((\mathsf{EVT}^* \times \mathbb{I}) \longrightarrow 2^{\mathsf{ASS}})$ associates to every possible world a repertoire function mapping a sequence of events and an agent to a set of assignments;
- $N : W \longrightarrow (\mathsf{EVT}^* \longrightarrow \mathsf{EVT})$ associates to every possible world a protocol function mapping sequences of events to events;
- $V : W \longrightarrow (\mathbb{P} \longrightarrow \{\mathtt{tt}, \mathtt{ff}\})$ associates to every possible world a valuation mapping propositional variables to truth values.

It is convenient to write $B_w(i)$, R_w, N_w, and V_w instead of $B(w)(i)$, $R(w)$, $N(w)$, and $V(w)$.

We require that B satisfies the following constraint, for all agents i and possible worlds w, w_1, w_2:

$$\text{if } w_1, w_2 \in B_w(i) \text{ then } B_{w_1}(i) = B_{w_2}(i) \tag{1.1}$$

We moreover require that R and N satisfy the following constraint, for all possible worlds w and sequences of events μ:

$$\text{if } N_w(\mu) = \delta \text{ then for every } i \in \mathbb{I}, \delta_i \in R_w(\mu, i) \tag{1.2}$$

The function B models the agents' uncertainty: $B_w(i)$ is the set of possible worlds that i cannot distinguish from the actual world w. Constraint 1.1 is nothing but transitivity and Euclideanity of the accessibility relations corresponding to the function B that are defined as $B_i = \{(w, w') \ : \ w' \in B_w(i)\}$.[2] It will make principles of positive and negative introspection valid. The tuple (W, B, V) is therefore a model of the standard logic of belief $\mathsf{K45}$. The set of assignments $R_w(\mu, i)$ models agent i's control after the sequence of event $\mu \in \mathsf{EVT}^*$ took place at w. The event $N_w(\mu)$ is the event that is going to happen after the sequence of events μ took place at w. For every possible world w, the 'next state' function maps sequences of events $\mu \in \mathsf{EVT}^*$ to events $N_w(\mu)$: if the sequence μ occurs then $N_w(\mu)$ is the next event that is going to happen afterwards. Constraint 1.2 is therefore a 'do implies can' principle. Note that it implies that for every i, $R_w(\mu, i)$ is non-empty. Finally, the valuation function V associates to every possible world a valuation: a mapping from the set of propositional variables to the set $\{\mathtt{tt}, \mathtt{ff}\}$.

For the subsequent definitions we need a bit of notation for the event sequences of EVT^*: we note nil the empty sequence of events, and we note $\delta;\mu$ the sequential composition of the event δ with the sequence of events μ.

Then the *update* of M by δ is defined as $M^\delta = (W, B, R^\delta, N^\delta, V^\delta)$, where R^δ, N^δ and V^δ are defined by:

$$\begin{array}{lll} R^\delta_w(\mu, i) & = R_w((\delta;\mu), i), & \text{for } \mu \in \mathsf{EVT}^* \text{ and } i \in \mathbb{I} \\ N^\delta_w(\mu) & = N_w(\delta;\mu), & \text{for } \mu \in \mathsf{EVT}^* \\ V^\delta_w(p) & = \begin{cases} \mathtt{tt} & \text{if } \exists i, \delta_i = p{\leftarrow}\top \text{ and } \nexists j, \delta_j = p{\leftarrow}\bot \\ \mathtt{ff} & \text{if } \exists i, \delta_i = p{\leftarrow}\bot \text{ and } \nexists j, \delta_j = p{\leftarrow}\top \\ V_w(p) & \text{otherwise} \end{cases} \end{array}$$

Hence updates neither change the set of possible worlds nor the agents' possibilities (the belief accessibility relation). Both the repertoire function and the protocol function are incremented by δ. The valuation function after the update gives to the variables that are assigned by δ their new truth values according to δ and leaves the other variables unchanged. The only subtlety is that we have to deal with contradicting assignments, i.e. when $\delta_i = p{\leftarrow}\top$ and $\delta_j = p{\leftarrow}\bot$, for some i and j. We here suppose that such contradicting assignments have no effect on p: the valuation $V^\delta_w(p)$ after the update is identical to the valuation before the update.

[2] The relation B_i is transitive if and only if $(w, w') \in B_i$ and $(w', w'') \in B_i$ implies $(w, w'') \in B_i$. It is Euclidean if and only if $(w, w') \in B_i$ and $(w, w'') \in B_i$ implies $(w', w'') \in B_i$.

The truth conditions are standard for $\top$, $\bot$, negation and disjunction, plus:

$$\begin{array}{ll} M, w \Vdash p & \text{iff } Val_w(p) = \mathtt{tt} \\ M, w \Vdash \mathsf{Hpn}(\delta_i) & \text{iff } (N_w(\mathsf{nil}))_i = \delta_i \\ M, w \Vdash \mathsf{Bel}_i\, \varphi & \text{iff } M, w' \Vdash \varphi \text{ for every } w' \in B_w(i) \\ M, w \Vdash \mathsf{Next}\, \varphi & \text{iff } M^{N_w(\mathsf{nil})}, w \Vdash \varphi \\ M, w \Vdash \langle\delta\rangle\varphi & \text{iff } \delta_i \in R_w(\mathsf{nil}, i) \text{ for all } i \in \mathbb{I} \text{ and } M^{\delta}, w \Vdash \varphi \end{array}$$

We say that a formula φ is *valid in the model* M, noted $M \Vdash \varphi$, if and only if $M, w \Vdash \varphi$ for every possible world w in M. The formula φ is *valid in the class of epistemic PC models* if and only if $M \Vdash \varphi$ for every epistemic PC model M. Finally, the formula φ is a *(global) logical consequence* of the set of formulas Γ, noted $\Gamma \models \varphi$, if and only if for every epistemic PC model M, if $M \Vdash \psi$ for every $\psi \in \Gamma$ then $M \Vdash \varphi$. Hence φ is valid if and only if $\emptyset \models \varphi$.

1.2.3 Reduction axioms and decidability

Basically, our semantics allows to eliminate the temporal operator Next and the dynamic operators $\langle\delta\rangle$. First, Next can be eliminated because the formula

$$\mathsf{Next}\,\varphi \leftrightarrow \bigvee_{\delta \in \mathsf{EVT}} \left(\langle\delta\rangle\varphi \wedge \bigwedge_{i \in \mathbb{I}} \mathsf{Hpn}(\delta_i) \right)$$

is valid. Note that finiteness of the set EVT warrants that the formula on the right is well-formed. Second, as customary in dynamic epistemic logics without the common belief operator, the dynamic operators $\langle\delta\rangle$ have reduction axioms.

Proposition 1. *The following equivalences are DDL−PA valid.*

$$\begin{array}{ll} \langle\delta\rangle\neg\varphi & \leftrightarrow \langle\delta\rangle\top \wedge \neg\langle\delta\rangle\varphi \\ \langle\delta\rangle(\varphi_1 \vee \varphi_2) & \leftrightarrow \langle\delta\rangle\varphi_1 \vee \langle\delta\rangle\varphi_2 \\ \langle\delta\rangle\mathsf{Bel}_i\, \varphi & \leftrightarrow \langle\delta\rangle\top \wedge \mathsf{Bel}_i\, [\delta]\varphi \end{array}$$

The above equivalences allow to 'push inwards' the modal operators $\langle\delta\rangle$ until they are no longer in the scope of the Boolean and the belief operators.

Proposition 2. *The equivalence*

$$\langle\delta\rangle p \leftrightarrow \begin{cases} \langle\delta\rangle\top & \text{if } \exists i \in \mathbb{I}, \delta_i = p{\leftarrow}\top \text{ and } \nexists j \in \mathbb{I}, \delta_j = p{\leftarrow}\bot \\ \bot & \text{if } \exists i \in \mathbb{I}, \delta_i = p{\leftarrow}\bot \text{ and } \nexists j \in \mathbb{I}, \delta_j = p{\leftarrow}\top \\ \langle\delta\rangle\top \wedge p & \text{else} \end{cases}$$

is DDL−PA valid.

In order to formally define the set of resulting formulas we recursively define *dynamic modalities*, noted μ, as abbreviations of sequences of dynamic operators:

$$\begin{array}{l} \langle\mathsf{nil}\rangle\varphi \stackrel{\mathtt{def}}{=} \varphi \\ \langle\mu;\delta\rangle\varphi \stackrel{\mathtt{def}}{=} \langle\mu\rangle\langle\delta\rangle\varphi \end{array}$$

Let us say that a *dynamic atom* is either a propositional variable from $\mathbb{P}$, or of the form $\langle\mu\rangle\top$, or of the form $\langle\mu\rangle\mathsf{Hpn}(\delta_i)$, where μ is a dynamic modality. The above propositions allows us to transform every DDL−PA formula φ into an equivalent formula in *reduced form*: a formula that is built from dynamic atoms by means of the Boolean and the Bel_i operators. Let $red(\varphi)$ be the *reduction* of φ that is obtained in this way.

Proposition 3. *The equivalence $\varphi \leftrightarrow red(\varphi)$ is DDL−PA valid.*

Proof. The result follows from Proposition 1, Proposition 2, and the fact that the rule of replacement of equivalents preserves validity. ■

If we consider dynamic atoms as propositional variables then every formula in reduced form is a formula of the logic of belief K45. This will be exploited now in order to give a decision procedure for our logic.

For every formula φ that is in reduced form, let $DM(\varphi)$ be the set of dynamic modalities of φ, i.e. the set of event sequences μ such that φ contains a dynamic atom of the form either $\langle\mu\rangle\top$, or $\langle\mu\rangle\mathsf{Hpn}(\delta_i)$ for some δ_i. Let Γ_φ be the set of formulas defined as follows.

$$\begin{array}{l}\Gamma_\varphi = \Big\{ \neg\Big(\langle\mu\rangle\mathsf{Hpn}(\delta_i) \wedge \langle\mu\rangle\mathsf{Hpn}(\delta_i')\Big) \ : \ \mu \in DM(\varphi), \delta \in \mathsf{EVT}, i \in \mathbb{I}, \text{ and } \delta_i \neq \delta_i' \Big\} \cup \\ \quad \{ \big(\bigwedge_{i\in\mathbb{I}} \langle\mu\rangle\mathsf{Hpn}(\delta_i)\big) \rightarrow \langle\mu\rangle\langle\delta\rangle\top \ : \ \mu \in DM(\varphi), \delta \in \mathsf{EVT}\} \cup \\ \quad \{ \langle\mu\rangle \big(\bigvee_{\delta\in\mathsf{EVT}} \bigwedge_{i\in\mathbb{I}} \mathsf{Hpn}(\delta_i)\big) \ : \ \mu \in DM(\varphi) \}\end{array}$$

Γ_φ axiomatises the properties of models and the semantic constraints that are relevant for φ: the first line says that N is a function, and the last line says that it is total.[3] The second line expresses the 'do implies can' constraint 1.2. The formula $\langle\mu\rangle\langle\delta\rangle\top$ has to be understood as a dynamic atom (i.e. it is of the form $\langle\mu'\rangle\top$ for an appropriate μ'). Observe that every Γ_φ is finite.

Proposition 4. *Let φ be a DDL−PA formula in reduced form. Then*

$$\varphi \text{ is DDL−PA valid in epistemic PC models if and only if } \Gamma_\varphi \models_{K45} \varphi,$$

where $\models_{K45}$ is the global consequence relation of the modal logic K45 and where in the latter check each dynamic atom is considered to be an atomic formula of the language of K45.

Proof. Let φ be a DDL−PA formula in reduced form. Define the set DA_φ of dynamic atoms that are relevant for φ as follows:

$$DA_\varphi = \mathbb{P} \cup \{\langle\mu\rangle\top \ : \ \mu \in DM(\varphi)\} \cup \{\langle\mu\rangle\mathsf{Hpn}(\delta_i) \ : \ \mu \in DM(\varphi), \delta \in \mathsf{EVT}, i \in \mathbb{I}\}$$

Suppose φ is invalid, i.e. $M, w \not\models \varphi$ for some epistemic PC-model

$$M = (W, B, R, N, V)$$

and possible world $w \in W$. We are going to build a K45 model

$$M^{\mathsf{K45}} = (W, B, V^{\mathsf{K45}})$$

[3] Note that the last constraint in Γ_φ can be dropped if $\mathbb{P}_\varphi \neq \mathbb{P}$: then it is not necessary to reflect the fact that the 'next' function N is total because then K45 models can always be arranged such that this is the case (by setting N_w to some δ such that say $\delta_i = p{\leftarrow}\top$ for some variable p not occurring in φ).

such that $M^{\mathsf{K45}} \models \psi$ for every $\psi \in \Gamma_\varphi$ and such that $M^{\mathsf{K45}} \not\models_{\mathsf{K45}} \varphi$, where the valuation V^{K45} of M^{K45} is for the set of propositional variables DA_φ. We define V^{K45} such that $V_w^{\mathsf{K45}}(\pi) = \mathtt{tt}$ iff $M, w \Vdash \pi$, for every dynamic atom $\pi \in DA_\varphi$. M^{K45} is a legal $\mathsf{K45}$ model because B is transitive and Euclidean. We check that $M, v \Vdash \psi$ for every $\psi \in \Gamma_\varphi$:

1. $M, v \Vdash \neg\left(\langle\mu\rangle\mathsf{Hpn}(\delta_i) \wedge \langle\mu\rangle\mathsf{Hpn}(\delta_i')\right)$ for every $\mu \in DM(\varphi)$, $\delta \in \mathsf{EVT}$, and $i \in \mathbb{I}$ such that $\delta_i \neq \delta_i'$ because N is a function.
2. $M, v \Vdash (\bigwedge_{i\in\mathbb{I}} \langle\mu\rangle\mathsf{Hpn}(\delta_i)) \to \langle\mu\rangle\langle\delta\rangle\top$ for every $\mu \in DM(\varphi)$ and $\delta \in \mathsf{EVT}$ because M satisfies the semantic constraint 1.2.
3. $M, v \Vdash \langle\mu\rangle(\bigvee_{\delta\in\mathsf{EVT}} \bigwedge_{i\in\mathbb{I}} \mathsf{Hpn}(\delta_i))$ for every $\mu \in DM(\varphi)$ because the function N is total.

We finally prove by induction on the form of ψ that for every possible world $v \in W$ and for every formula ψ we have $M, v \vdash \psi$ iff $M^{\mathsf{K45}}, v \vdash \psi$. The base case of the induction where ψ is a dynamic atom is clear by the definition of V^{K45}; and the induction step is obvious.

Suppose $\Gamma_\varphi \not\models_{\mathsf{K45}} \varphi$, i.e. there is a $\mathsf{K45}$ model

$$M^{\mathsf{K45}} = (W, B, V^{\mathsf{K45}})$$

such that $M^{\mathsf{K45}} \models \psi$ for every $\psi \in \Gamma_\varphi$ and such that $M^{\mathsf{K45}}, w \not\models_{\mathsf{K45}} \varphi$ for some $w \in W$ (where the valuation V^{K45} of M^{K45} is for the set of propositional variables DA_φ). We build an epistemic PC model

$$M = (W, B, R, N, V)$$

by defining R, N, and V as follows:

$$R_w(\mu, i) = \begin{cases} \{\delta_i \ : \ V_w^{\mathsf{K45}}(\langle\mu\rangle\langle\delta\rangle\top) = \mathtt{tt}\} & \text{if } \mu \in DM(\varphi) \\ \{\delta_i^0\} & \text{if } \mu \notin DM(\varphi) \end{cases}$$

$$N_w(\mu) = \begin{cases} \delta & \text{if } \mu \in DM(\varphi) \text{ and for every } i \in \mathbb{I}_\varphi, V_w^{\mathsf{K45}}(\langle\mu\rangle\mathsf{Hpn}(\delta_i)) = \mathtt{tt} \\ \delta^0 & \text{else} \end{cases}$$

$$V_w(p) = V_w^{\mathsf{K45}}(p) \text{ for } p \in \mathbb{P}$$

In the definition of R and N, δ^0 is an arbitrary fixed event from EVT. Observe that N is well defined: it is a function because of the first item of the definition of Γ_φ, and it is total because of the third item. Moreover, M satisfies constraint 1.2 because of the second item of the definition of Γ_φ and because we have enforced it when $\mu \notin DM(\varphi)$. Finally, M's accessibility relation B is transitive and Euclidean because the relation B of M^{K45} is so. It remains to establish that $M, w \nVdash \varphi$. We do so by proving inductively that for every formula ψ such that $DM(\psi) \subseteq DM(\varphi)$ and for every possible world $v \in W$ we have $M, v \Vdash \psi$ if and only if $M^{\mathsf{K45}}, v \Vdash \psi$. It follows that φ is invalid in epistemic PC models. ■

It follows from Proposition 3 and Proposition 4 that checking validity of φ in our logic reduces to checking whether $red(\varphi)$ is a global consequence of Γ_φ in $\mathsf{K45}$. Furthermore, it follows from the decidability of the $\mathsf{K45}$ global consequence relation that $\mathsf{DDL{-}PA}$ validity is decidable.

Theorem 1. *The DDL−PA validity problem is decidable.*

Proof. Let φ be a $\mathsf{DDL{-}PA}$ formula. By Proposition 2 φ is valid iff $red(\varphi)$ is valid. As the latter is in reduced form, by Proposition 4 it is valid iff $\Gamma_\varphi \models_{\mathsf{K45}} \varphi$, i.e. iff φ

is a global logical consequence of Γ_φ in $\mathsf{K45}$. As the latter problem is decidable, it follows that the problem of validity in epistemic PC models is decidable. ■

The reduction schemas of Proposition 1 and Proposition 2 might increase formula size exponentially. Reduction therefore only provides a suboptimal decision procedure, and it remains to establish the complexity of validity checking.

1.2.4 Axiomatisation

An axiomatisation of $\mathsf{DDL{-}PA}$ can be obtained by putting together:

- some axiomatisation of multiagent $\mathsf{K45}$
- the schemas of Proposition 1 and Proposition 2
- the axiom schemas:

$$\begin{array}{ll} \neg(\mathsf{Hpn}(\delta_i) \wedge \mathsf{Hpn}(\delta_i')) , & \text{for } \delta_i \neq \delta_i' \\ \mathsf{Hpn}(\delta_i) \rightarrow \langle \delta_i \rangle \top & \\ \bigvee_{\delta \in \mathsf{EVT}} \bigwedge_{i \in \mathbb{I}} \mathsf{Hpn}(\delta_i) & \end{array}$$

- the inference rule:

$$\text{from } \varphi \leftrightarrow \psi \text{ infer } \langle \delta \rangle \varphi \leftrightarrow \langle \delta \rangle \psi$$

The above three axiom schemas generalise the formulas in Γ_φ of our reduction from $\mathsf{DDL{-}PA}$ to $\mathsf{K45}$.

1.3 Defining trust in DDL−PA

By now we have the ingredients we need in order to capture Castelfranchi and Falcone's relevant concepts. We have already seen that our logic directly accounts for belief. It remains to define the concepts of goal, willingness, ability and opportunity.

But first of all let us have a closer look at the concept of belief.

1.3.1 Belief and probability

In epistemic logic, "i believes that φ" is identified with "φ is true in all worlds that are possible for i". This is a strong form of belief. We also need a weaker form where belief is identified with truth in the most plausible possible worlds. We therefore have:

- the strong belief in some physical properties of the world, which can be objectively verified, such as an agent's own beliefs and intentions or the belief that some agent is able to perform some action, or that some action will achieve some outcome;
- the weak belief in some non verifiable properties such as another agent's beliefs and intentions.

We are going to model an agent's weak beliefs in terms of a subset of the set of worlds possible for i.

We suppose that $\mathbb{P}$ contains special propositional variables poss, that we read "the current state is plausible/probable". Note that poss being a propositional variable, its truth value can be changed: the events $\mathsf{poss}{\leftarrow}\top$ and $\mathsf{poss}{\leftarrow}\bot$ modify the evaluation of the current state (whether the current state is plausible/probable or not). We then introduce modal operators of *probability* Prob_i as follows:

$$\mathsf{Prob}_i\,\varphi \stackrel{\mathtt{def}}{=} \mathsf{Bel}_i\,(\mathsf{poss} \to \varphi)$$

We therefore identify "it is probable for agent i that φ" with "φ is true in all worlds that i considers plausible/probable". Our definition leads to both positive and negative introspection principles for weak belief: the formula schemas $\mathsf{Prob}_i\,\varphi \to \mathsf{Bel}_i\,\mathsf{Prob}_i\,\varphi$ and $\neg\mathsf{Prob}_i\,\varphi \to \mathsf{Bel}_i\,\neg\mathsf{Prob}_i\,\varphi$ are both valid. It also leads to the natural principle $\mathsf{Bel}_i\,\varphi \to \mathsf{Prob}_i\,\varphi$.

1.3.2 Goal and Willingness

In the same vein as before, we suppose that $\mathbb{P}$ contains special propositional variables $\mathsf{good}(i)$, one for every $i \in \mathbb{I}$. $\mathsf{good}(i)$ reads "the current state is good for i". Note that i's evaluation of the current state can be modified by assignments.

We introduce preference operators Pref_i that are defined as follows:

$$\mathsf{Pref}_i\,\varphi \stackrel{\mathtt{def}}{=} \mathsf{Bel}_i\,(\mathsf{good}(i) \to \varphi)$$

We therefore identify "agent i prefers that φ" with "φ is true in all of i's possible worlds that are good for i". Our definition leads to positive and negative introspection principles for preferences: the formula schemas $\mathsf{Pref}_i\,\varphi \to \mathsf{Bel}_i\,\mathsf{Pref}_i\,\varphi$, and $\neg\mathsf{Pref}_i\,\varphi \to \mathsf{Bel}_i\,\neg\mathsf{Pref}_i\,\varphi$ are both valid. Moreover the principle of strong realism $\mathsf{Bel}_i\,\varphi \to \mathsf{Pref}_i\,\varphi$ is valid [3, 4].

We identify j's *willingness* to perform δ_j with j's preference that δ_j happen next, formally: $\mathsf{Pref}_j\,\mathsf{Hpn}(\delta_j)$.

We finally identify "agent i has goal that φ" with "i prefers that $\mathsf{Next}\,\varphi$ is true", formally: $\mathsf{Pref}_i\,\mathsf{Next}\,\varphi$.

1.3.3 Ability and opportunity

We start by defining the concept "group J is capable to achieve φ by doing δ_J, if the agents outside J cooperate".

$$\langle\delta_J\rangle\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\delta'_{-J}} \langle\delta_J{\cdot}\delta'_{-J}\rangle\varphi$$

where $\delta_J{\cdot}\delta'_{-J}$ is the function that is obtained in the obvious way by combining the functions δ_J and δ'_{-J} (whose domains are disjoint):

$$(\delta_J{\cdot}\delta'_{-J})(i) \;=\; \begin{cases}\delta_J(i) & \text{if } i \in J\\ \delta'_{-J}(i) & \text{if } i \notin J\end{cases}$$

Observe that $\langle\delta_{\mathbb{I}}\rangle\varphi = \langle\delta\rangle\varphi$.

We define the *ability* of a group J to perform a joint action δ_J as $\langle\delta_J\rangle\top$.

Opportunity of J to achieve φ by doing δ_J is defined by means of the dual modal operator as: $[\delta_J]\varphi \stackrel{\texttt{def}}{=} \neg\langle\delta_J\rangle\neg\varphi$.

1.3.4 Trust

Now we are in a position to define the trust predicate $\mathsf{Trust}(i, J, \delta_J, \varphi)$, read "agent i trusts group J to do δ_J in order to achieve i's goal φ", as the conjunction of the following four formulas:

1. $\mathsf{Pref}_i\,\mathsf{Next}\,\varphi$: agent i has the goal that φ;
2. $\mathsf{Bel}_i\,\langle\delta_J\rangle\top$: i believes J is able to perform δ_J;
3. $\mathsf{Prob}_i\,\bigwedge_{j\in J}\mathsf{Pref}_j\,\mathsf{Hpn}(\delta_j)$: i expects the members of J to be willing to perform their part δ_j;
4. $\mathsf{Bel}_i\,[\delta_J]\varphi$: i believes J has the opportunity to achieve φ by performing δ_J, no matter what other agents do.

Actually we might drop the second argument J from the trust predicate because the third argument δ_J contains already that information.

Proposition 5. *The implication*

$$\bigwedge_{j\in J}\mathsf{Trust}(i, j, \delta_j, \varphi) \to \mathsf{Trust}(i, J, \delta_J, \varphi)$$

is DDL−PA valid.

Note that the converse does not hold. (This is due to the fourth item in the definition of trust.)

Proposition 6. *Let $J_1 \cap J_2 = \emptyset$. The implication*

$$\mathsf{Trust}(i, J_1, \delta_{J_1}, \varphi) \wedge \mathsf{Trust}(i, J_2, \delta_{J_2}, \psi) \to \mathsf{Trust}(i, J_1 \cup J_2, \delta_{J_1}{\cdot}\delta_{J_2}, \varphi \wedge \psi)$$

is DDL−PA valid.

1.4 A more general definition of trust

The evaluation of the trust predicate involves universal quantification over the complement set $-J = \mathbb{I} \setminus J$. If the truster i is not in J then that quantification includes all

possible actions that i may perform. This is not realistic. More generally, we might wish to fix the behaviour of some of the agents outside J. The following definition achieves this, moving to a 5-ary predicate $\mathsf{Trust}(i, J, \delta_J, \varphi, \delta_K)$, where J and K are disjoint subsets of $\mathbb{I}$. It reads "i trusts J to do δ_J in order to achieve i's goal φ, given that the agents of K perform δ_K".

It is only the opportunity condition of the previous definition that has to be redefined:

4′. $\mathsf{Bel}_i\,[\delta_J \cdot \delta_K]\varphi$: i believes J has the opportunity to achieve φ by performing δ_J, given that the agents in K do δ_K, and no matter what other agents outside of J and K do.

Our previous definition of trust is a particular case of the new definition.

Proposition 7. *The equivalence*

$$\mathsf{Trust}(i, J, \delta_J, \varphi) \leftrightarrow \mathsf{Trust}(i, \delta_J, \varphi, \delta_\emptyset)$$

is DDL−PA valid.

1.5 Why we need strong and weak beliefs

A natural requirement to be added to our logic is the following axiom of intentional action:

$$\Big(\langle \delta_j \rangle \top \wedge \mathsf{Pref}_j\, \mathsf{Hpn}(\delta_j)\Big) \rightarrow \mathsf{Hpn}(\delta_j)$$

It follows that when i trusts j to perform δ_j then i believes that j is actually going to perform δ_j.

The above requirement allows us to give a technical explanation why we need two different kinds of belief. Let us suppose that the willingness condition is not

$$\mathsf{Prob}_i \bigwedge_{j \in J} \mathsf{Pref}_j\, \mathsf{Hpn}(\delta_j),$$

but $\mathsf{Bel}_i \bigwedge_{j \in J} \mathsf{Pref}_j\, \mathsf{Hpn}(\delta_j)$. With the ability condition $\mathsf{Bel}_i \langle \delta_J \rangle \top$ it follows that

$$\mathsf{Bel}_i \bigwedge_{j \in J} \mathsf{Hpn}(\delta_j).$$

Then with the opportunity condition $\mathsf{Bel}_i\,[\delta_J]\varphi$ it follows that $\mathsf{Bel}_i\,\mathsf{Next}\,\varphi$. But the latter implies the goal condition $\mathsf{Pref}_i\,\mathsf{Next}\,\varphi$, which would therefore be redundant with the goal condition of the above definition of trust!

We therefore chose to weaken the willingness condition.

1.6 Conclusion

We have defined a simple modal logic of belief and action DDL−PA. The actions of DDL−PA are assignments of propositional variables that are simultaneously performed by the agents. We have shown that our logic is decidable.

Our logic captures the spirit of Castelfranchi and Falcone's theory of trust. We have argued that a distinction between strong and weak belief is needed in order to keep the four items defining trust independent.

References

1. C. Castelfranchi and R. Falcone. 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, 1998.
2. C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley and Sons, Chichester, UK, 2010.
3. Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42(2–3):213–261, 1990.
4. Philip R. Cohen and Hector J. Levesque. Persistence, intentions, and commitment. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, chapter 3, pages 33–69. MIT Press, Cambridge, MA, 1990.
5. Hans P. van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Kluwer Academic Publishers, 2007.
6. R. Falcone and C. Castelfranchi. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
7. A. Herzig, E. Lorini, J. F. Hübner, and L. Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244, 2010.
8. A. Herzig, E. Lorini, F. Moisan, and N. Troquard. A dynamic logic of normative systems. In *Proceedings of IJCAI 2011*, 2011. to appear.
9. Wiebe van der Hoek, Dirk Walther, and Michael Wooldridge. On the logic of cooperation and the transfer of control. *J. of AI Research (JAIR)*, 37:437–477, 2010.
10. Wiebe van der Hoek and Michael Wooldridge. On the logic of cooperation and propositional control. *Artif. Intell.*, 164(1-2):81–119, 2005.