A logic of trust and reputation

Andreas Herzig, Emiliano Lorini
IRIT, Toulouse, France

Jomi F. Hübner, Laurent Vercouter
ENS Mines Saint-Etienne, France

Abstract

The aim of this paper is to present a logical framework in which the concepts of trust and reputation can be formally characterized and their properties studied. We start from the definition of trust proposed by Castelfranchi & Falcone (C&F). We formalize this definition in a logic of time, action, beliefs and choices. Then, we provide a refinement of C&F's definition by distinguishing two general types of trust: occurrent trust and dispositional trust. In the second part of the paper we present a definition of reputation that is structurally similar to the definition of trust but moves the basic concept of belief to a collective dimension of group belief.

Contents

1 Introduction

Several disciplines in the area of social science such as economics [51, 1], sociology [14, 42] and social philosophy [3, 33] have been interested in the concepts of trust and reputation. For instance, when looking at human organizations, social scientists have been mostly interested in individualizing the antecedents of collective behavior and collective action between interacting individuals and roles. A central concern of the field has been to identify the determinants of cooperation, coordination and delegation [51, 1]. Among the different determinants, trust and reputation have been recognized as two of the most important ones [14, 42].

The concepts of trust and reputation are particularly important in domains where agent technologies are applied, such as information retrieval, e-commerce, and more generally in peer-to-peer systems. They have been in the focus of many research projects in the last couple of years, and now there exist manifold theoretical models and implemented systems in the multi-agent system domain [60, 56]. One of the most prominent theoretical models is the cognitive theory of trust by Castelfranchi and Falcone, henceforth abbreviated C&F theory [24, 11]. Their definition of trust is formulated as an individual belief of the truster about some properties of the trustee that are relevant for the achievement of a given goal. These properties include ascription of mental attitudes, abilities, and opportunities.

Our first aim in this article is to give a formal and more refined version of the C&F definition. We adopt their reduction of trust to the more primitive concepts of belief, goal, capability and opportunity. We then formalize the concept of trust in a logic of time, action, beliefs and chosen goal. Our refinement of the C&F theory of trust consists in distinguishing two general types of trust: occurrent trust and dispositional trust. The former type of trust is the one studied by Castelfranchi and Falcone, and it is the trust in the occurrence of an action α of the trustee *here* and *now* with respect to a current goal of the truster. The latter type of trust is the truster's belief that it will possibly have a certain goal φ in the future and, whenever it will have such a goal and certain conditions obtain, the trustee will perform α and thereby will ensure φ. In this sense, the truster is disposed to trust the trustee. An example of this kind of trust is dispositional trust in the context of trade: the truster i believes it to be possible that in the future it will have the goal to receive a certain product from j, and believes that whenever it will have this goal and will pay the product to j, then j will send the product to i so that it will receive the product.

Our second aim in this article is to propose a definition of reputation that is structurally similar to the definition of trust but moves the basic concept of belief to a collective dimension of group belief. In fact, in our perspective, trust and reputation have the same content: the properties of a given target (dispositions and capabilities) which are relevant for the accomplishment of a given task. The only difference is that trust is an individual attitude (micro level), whereas reputation is a group attitude (macro level). In particular, while trust is an evaluation of a given target j by a certain individual agent i, reputation is a collective evaluation of a given target j by a group of agents I.

It is to be noted that the analysis of trust and reputation that we propose in this article is qualitative. We are interested here in identifying the basic constituents of trust and reputation, and in explaining the logical properties of the latter on the basis

of the logical properties of the former. It is beyond the objectives of the present work to study the quantitative dimension of trust (i.e. *graded trust*) in terms of probability theory or some statistical methods.

The present article is organized as follows. After recalling the C&F definition and informally distinguishing occurrent from dispositional trust in Section 2, we introduce a logical framework which allows to reason about time, actions, beliefs and chosen goals (Section 3). In Section 4, we give a formal logical analysis of occurrent trust and dispositional trust. We also compare our definitions of trust with alternative definitions of trust which have been proposed in philosophy and economics. In Section 5, we move from trust to reputation: first, we extend the logic of Section 3 with modal operators which allow to express a kind of group belief, viz. that a certain fact φ is public for a group I; then we propose a definition of reputation in terms of a group belief about some properties of a target which parallels that of dispositional trust. In the last part of the article (Section 6), we evaluate to which extent the existing computational models of trust and reputation proposed in the multi-agent systems (MAS) domain conform to our formal definitions of trust and reputation.

2 The concept of trust: an informal view

Differently from other more computationally oriented approaches [39, 66], in C&F theory trust is not reduced to mere subjective probability that is updated in the light of direct interaction with the trustee and reputational information. The C&F theory of social trust accounts for the truster's *attribution process*, that is, for the truster's ascription of properties to the trustee (capabilities, intention, dispositions, etc.) and to the environment in which the trustee is going to act, which are together sufficient to ensure that one of the truster's goals will be achieved. In this perspective, trust is nothing more than the truster's *evaluation* of certain relevant properties of the trustee. It is to be noted that C&F also consider a related notion of *decision to trust*, that is, the truster's decision to bet and wager on the trustee and to rely on it for the accomplishment of a given task (on the distinction between trust as an *evaluation* and trust as a *decision*, see also [24, 52]). This second type of trust will not be studied in the present article.

Two fundamental distinctions are introduced in C&F theory:

- between a dimension of internal attribution of trust (i's trust *in the 'good will' of* j) and a dimension of external attribution of trust (the environment trust: i's trust in the environment about the effects of j's action);
- between the different dimensions of the truster's evaluation of and expectation about the trustee's properties, in particular its quality (that is due to his skills and capabilities), and the expectation about the certainty of the expected/desired behavior of the trustee (i.e. the truster's expectation that the trustee is willing to act in a certain way).

According to C&F theory, trust has four ingredients: a truster i, a trustee j, an action α of j, and a goal φ of i.[1] In their definition, "i trusts j to do α with respect to φ" if and

[1] Throughout the paper we use α to denote actions and φ to denote goals.

only if:

1. i has the goal φ;
2. i believes that
   (a) j is capable to do α;
   (b) j, by doing α, will ensure φ;[2] and
   (c) j intends to do α.

For example, when i trusts j to send product P in view of satisfying i's goal of possessing P then (1) i wants to possess P, (2a) i believes that j is capable to send P, (2b) that j's sending P will result in i possessing P, and (2c) that j has the intention to send P.

Castelfranchi and Falcone stress the importance of the goal component in the definition of trust (condition 1). Indeed, i trusts j to do α only if α is relevant for i's goals. This condition allows to distinguish trust from mere *thinking* and *foreseeing*, and is the main difference with Jones's analysis of what he calls 'core trust' [37, 38]. In C&F theory, condition (2a) and condition (2b) relate to the external attribution of i's trust, while the intention concept (condition (2c)) relates to the internal attribution.

One of the objectives of this article is to refine C&F theory by distinguishing two kinds of trust: in the first case, the truster believes that the trustee is going to act *here and now*; in the second, the truster believes that the trustee is going to act *whenever some conditions are satisfied*. This relates to a standard distinction in philosophy of action: *action tokens* (alias concrete actions, or action occurrences) are unique, e.g. agent j's selling of good P at time t; *action types* are repeatable, e.g. j's action of turning the head, or of paying. Action tokens are instances of action types [28]. Therefore in the first case we are going to use the term *occurrent trust*: trust in occurrence of action instance α here and now. In the second case, we are going to use the term *dispositional trust*: trust in a general disposition of the trustee to perform an instance of the action type α. This distinction between occurrent trust and dispositional trust relates to distinction between occurrent belief and dispositional belief employed by some philosophers (e.g. [61]).

C&F theory is only about occurrent trust: when i trusts j's action (instance) of sending i some product P, then i believes that j's next action is to send P. The trustee's actual intention to perform α together with its capability to perform α logically entail that it is indeed going to perform α. Suppose that i currently does not have the goal to possess product P. According to the C&F definition of occurrent trust, it is not the case that i actually trusts j about sending product P. It seems nevertheless natural to allow for some kind of trust in this case, e.g. when possessing P is a potential goal for i, and when i believes that j will send P whenever i will have the goal to possess the product and has paid for it.

[2] Note that C&F theory often refers to the second property (2b) of the truster as j's opportunity to ensure φ by doing α,or j's power to ensure φ by doing α.

Generally speaking, when i trusts j's action type α with respect to φ in the circumstances κ (dispositional trust), then i believes that, whenever it will have the goal that φ and the conditions κ obtain, then j will ensure φ by doing α.

The concepts of occurrent trust and dispositional trust will be formally characterized in Section 4. We will show that the goal component in dispositional trust is logically weaker than that of occurrent trust: the latter implies the former, but not the other way round. Moreover, we will show that it is possible to infer occurrent trust from dispositional trust, under some conditions. But before developing our formal analysis of trust, we need to set up our logical language. This is what we will do in the next section.

3 A logic for trust

We present in this section the multimodal logic $\mathcal{L}$ that will be used in Section 4 to formalize the concepts of occurrent and dispositional trust. $\mathcal{L}$ combines the expressiveness of dynamic logic [31] and temporal logic with the expressiveness of a logic of belief and choice that can be used to define intention (and may be called BDI-like, see [13] for instance). Our logic has two dynamic operators $\texttt{After}_{i:\alpha}$ and $\texttt{Does}_{i:\alpha}$. The first quantifies universally over all possible executions of action α (just as the dynamic logic action operator), while the second existentially quantifies over the actual execution of α, and allows to talk about the action actually occurring. It extends more standard logics of action and belief such as those of dynamic logic [31] which only have a single action operator. The modal operator $\texttt{After}_{i:\alpha}$ allows to talk about an agent's capabilities, while $\texttt{Does}_{i:\alpha}$ allows to talk about what an agent does and what he intends to do.

3.1 Syntax

The syntactic primitives of the logic $\mathcal{L}$ are the following:

- a nonempty finite set of individual agents $AGT = \{i, j, \ldots\}$;
- a nonempty finite set of atomic actions $ACT = \{\alpha, \beta, \ldots\}$;
- a nonempty set of atomic formulas $ATM = \{p, q, \ldots\}$.

The language of $\mathcal{L}$ is the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \texttt{G}\varphi \mid \texttt{After}_{i:\alpha}\varphi \mid \texttt{Does}_{i:\alpha}\varphi \mid \texttt{Bel}_i\varphi \mid \texttt{Choice}_i\varphi$$

where p ranges over ATM, α ranges over ACT and i ranges over AGT.

The operators of our logic have the following reading.

- $\texttt{G}\varphi$: 'φ will always be true'.
- $\texttt{After}_{i:\alpha}\varphi$: 'immediately after agent i does α, it is the case that φ' (therefore $\texttt{After}_{i:\alpha}\bot$ is read 'agent i cannot do action α'). In the sequel we will often read $\texttt{After}_{i:\alpha}\varphi$ as 'i has the opportunity to ensure φ by doing α'.

- $\texttt{Does}_{i:\alpha}\varphi$: 'agent i is going to do α and φ will be true afterward' (therefore $\texttt{Does}_{i:\alpha}\top$ is read: 'agent i is going to do α').
- $\texttt{Bel}_i\varphi$: 'agent i believes that φ'.
- $\texttt{Choice}_i\varphi$: 'agent i has the chosen goal that φ' (which can be shortened to 'agent i wants φ to be true').

$\texttt{G}$ ('globally') is a temporal modality which is used to express facts that are always true in the (strict) future.

Operators $\texttt{Choice}_i$ are used to denote an agent's current chosen goals, that is, the goals that the agent has decided to pursue here and now. We do not consider how an agent's chosen goals originate through deliberation from more primitive motivational attitudes called desires and from moral attitudes such as ideals and imperatives (see [57, 15, 12] on this issue). Since an agent's chosen goals result from the agent's deliberation, they must satisfy two fundamental rationality principles: chosen goals have to be consistent (i.e., a rational agent cannot decide to pursue an inconsistent state of affairs); chosen goals have to be compatible with the agent's beliefs (i.e., a rational agent cannot decide to pursue something that it believes to be impossible). These two principles will be formally expressed in Section 3.3.

The following abbreviations will be convenient:

$$\begin{aligned}
\texttt{G}^*\varphi &\stackrel{\text{def}}{=} \varphi \wedge \texttt{G}\varphi \\
\texttt{Capable}_i(\alpha) &\stackrel{\text{def}}{=} \neg\texttt{After}_{i:\alpha}\bot \\
\texttt{Intends}_i(\alpha) &\stackrel{\text{def}}{=} \texttt{Choice}_i\texttt{Does}_{i:\alpha}\top \\
\texttt{F}\varphi &\stackrel{\text{def}}{=} \neg\texttt{G}\neg\varphi \\
\texttt{F}^*\varphi &\stackrel{\text{def}}{=} \neg\texttt{G}^*\neg\varphi \\
\texttt{Poss}_i\varphi &\stackrel{\text{def}}{=} \neg\texttt{Bel}_i\neg\varphi
\end{aligned}$$

$\texttt{G}^*\varphi$ stands for 'φ is true in the present and will always be true'. $\texttt{Capable}_i(\alpha)$ stands for 'i has the capability to do α' (which can be shortened to 'agent i can do action α'). $\texttt{Intends}_i(\alpha)$ stands for 'agent i intends to do action α'. $\texttt{F}\varphi$ stands for 'φ will eventually be true'. Finally, $\texttt{Poss}_i\varphi$ stands for 'according to i, φ is possible' (or 'i does not exclude φ').

3.2 Semantics

We first define Kripke frames, and then models and truth conditions for the logical connectives.

3.2.1 Frames

Frames of the logic $\mathcal{L}$ ($\mathcal{L}$-frames) are tuples $F = \langle W, A, B, C, D, G\rangle$ defined as follows.

- W is a nonempty set of possible worlds or states.
- $A : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a relation $A_{i:\alpha}$ between possible worlds in W.

- $B : AGT \longrightarrow W \times W$ maps every agent i to a serial, transitive and Euclidean[3] relation B_i between possible worlds in W.
- $C : AGT \longrightarrow W \times W$ maps every agent i to a serial relation C_i between possible worlds in W.
- $D : AGT \times ACT \longrightarrow W \times W$ maps every agent i and action α to a deterministic relation $D_{i:\alpha}$ between possible worlds in W.[4]
- G is a transitive and connected[5] relation on W.

It is convenient to view relations on W as functions from W to 2^W; therefore we write $A_{i:\alpha}(w)$ for the set $\{w' : (w, w') \in A_{i:\alpha}\}$, etc. $B_i(w)$ is the set of worlds that are compatible with agent i's beliefs at w; and $C_i(w)$ is the set of worlds that are compatible with agent i's choices at w. Given a possible world $w' \in W$, $G(w)$ is the set of worlds w' that are in the future of w. $A_{i:\alpha}(w)$ is the set of worlds w' that can be reached from w through the occurrence of agent i's action α. If $(w, w') \in D_{i:\alpha}$ then w' is the unique actual *successor* world of w, that will be reached from w through the occurrence of agent i's action α at w. (We might also say that $D_{i:\alpha}$ is a partial function.) If $D_{i:\alpha}(w) \neq \emptyset$ (resp. $A_{i:\alpha}(w) \neq \emptyset$) then, we say that $D_{i:\alpha}$ (resp. $A_{i:\alpha}$) is defined at w. We therefore have two kinds of relations for specifying the dynamic dimension of frames:

- when $D_{i:\alpha}(w) = \{w'\}$ then at w agent i performs an action α, resulting in the next state w';
- when $w' \in A_{i:\alpha}(w)$ then *if* at w agent i performs α then this *might* result in w'.

Therefore, when $w' \in A_{i:\alpha}(w)$ but $D_{i:\alpha}(w) = \emptyset$ then at w agent i does not perform α, but if it did so it might have produced another outcome world w'.

Frames will have to satisfy some constraints in order to be legal $\mathcal{L}$-frames. For every $i, j \in AGT$, $\alpha, \beta \in ACT$ and $w \in W$ we suppose:

S1 if $D_{i:\alpha}$ and $D_{j:\beta}$ are defined at w then $D_{i:\alpha}(w) = D_{j:\beta}(w)$.

Constraint **S1** says that if w' is the *next* world of w which is reached from w through the occurrence of agent i's action α and w'' is also the *next* world of w which is reached from w through the occurrence of agent j's action β, then w' and w'' denote the same world. Indeed, we suppose that every world can only have one *next* world. Note that **S1** implies determinism of every $D_{i:\alpha}$.

Moreover, for every $i \in AGT$, $\alpha \in ACT$ we suppose:

S2 $D_{i:\alpha} \subseteq A_{i:\alpha}$.

[3] A relation B_i on W is Euclidean if and only if, if $(w, w') \in B_i$ and $(w, w'') \in B_i$ then $(w', w'') \in B_i$.

[4] A relation $D_{i:\alpha}$ is deterministic iff, if $(w, w') \in D_{i:\alpha}$ and $(w, w'') \in D_{i:\alpha}$ then $w' = w''$.

[5] The relation G is connected iff for every $w \in W$ we have: if $(w, w') \in G$ and $(w, w'') \in G$ then $(w', w'') \in G$ or $(w'', w') \in G$ or $w' = w''$.

The constraint **S2** says that if w' is the *next* world of w which is reached from w through the occurrence of agent i's action α, then w' must be a world which is *reachable* from w through the occurrence of agent i's action α.

The following semantic constraints **S3** and **S4** are about the relationship between an agent i's choices (i.e., chosen worlds) and the actions performed by i. For every $i \in AGT$, $\alpha \in ACT$ and $w \in W$, we suppose that:

S3 if $A_{i:\alpha}$ is defined at w and $D_{i:\alpha}$ is defined at w' for all $w' \in C_i(w)$ then $D_{i:\alpha}$ is defined at w;

S4 if $w' \in C_i(w)$ and $D_{i:\alpha}$ is defined at w, then $D_{i:\alpha}$ is defined at w'.

As to time and actions, we suppose that worlds resulting from an action α performed at world w are in the future of w. That is, for every $i \in AGT$, $\alpha \in ACT$:

S5 $D_{i:\alpha} \subseteq G$.

Moreover, we suppose an action cannot "jump" to a distant future world that is more than one time step away, i.e., if a world w' is accessible from world w via action α performed by i then, for every future world w'' of w, either w'' is in the future of w' or $w' = w''$. Formally, for every $\alpha \in ACT$, $i \in AGT$ and $w \in W$:

S6 if $w' \in D_{i:\alpha}(w)$ and $v \in G(w)$ then $w' = v$ or $v \in G(w')$.

This semantic constraint ensures that there is no third future world between a world w and the outcome w' of an action starting at w. This is the reason why $D_{i:\alpha}(w) = \{w'\}$ can be considered as the *next* state of w which is reached through the occurrence of action α performed by i.

The next constraint relates worlds that are compatible with agent i's beliefs and worlds that are compatible with i's chosen goals: as motivated in the beginning of Section 3.1, they should not be disjoint. For every $i \in AGT$ and $w \in W$:

S7 $C_i(w) \cap B_i(w) \neq \emptyset$.

The last constraint on $\mathcal{L}$-frames is one of introspection w.r.t. choices. For every $i \in AGT$ and $w \in W$:

S8 if $w' \in B_i(w)$ then $C_i(w) = C_i(w')$.

3.2.2 Models and truth conditions

Models of the logic $\mathcal{L}$ ($\mathcal{L}$-models) are tuples $M = \langle F, V \rangle$ defined as follows.

- F is a $\mathcal{L}$-frame.
- $V : W \longrightarrow 2^{ATM}$ is a truth assignment which associates each world w with the set $V(w)$ of atomic formulas true in w.

Given a model M, a world w and a formula φ, we write $M, w \models \varphi$ to mean that φ is true at world w in M. The rules defining the truth conditions of formulas are just standard for atomic formulas, negation and disjunction. The following are the remaining truth conditions for $\mathtt{G}\varphi$, $\mathtt{After}_{i:\alpha}\varphi$, $\mathtt{Does}_{i:\alpha}\varphi$, $\mathtt{Bel}_i\varphi$ and $\mathtt{Choice}_i\varphi$.

- $M, w \models \mathtt{G}\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in G$.
- $M, w \models \mathtt{After}_{i:\alpha}\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in A_{i:\alpha}$.
- $M, w \models \mathtt{Does}_{i:\alpha}\varphi$ iff there is $w' \in D_{i:\alpha}(w)$ such that $M, w' \models \varphi$.
- $M, w \models \mathtt{Bel}_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in B_i$.
- $M, w \models \mathtt{Choice}_i\varphi$ iff $M, w' \models \varphi$ for all w' such that $(w, w') \in C_i$.

Observe that the modal operator $\mathtt{Does}_{i:\alpha}$ is of type possibility, and that all other modal operators are of type necessity.

We write $\models_{\mathcal{L}} \varphi$ if φ is *valid* in all $\mathcal{L}$-models, i.e. $M, w \models \varphi$ for every $\mathcal{L}$-model M and world w in M. Finally, we say that φ is *satisfiable* if there exists a $\mathcal{L}$-model M and world w in M such that $M, w \models \varphi$.

3.3 Axiomatization

Figure 1 contains the axiomatization of the logic $\mathcal{L}$.

All our modal operators have the rule of necessitation (the last five inference rules). The first group of axioms is for the belief operator: its logic is KD45 [32]. The next group is for the choice operators, whose logic is KD. These operators are similar to Cohen & Levesque's goal operators [13]. Thus, we suppose positive and negative introspection for beliefs (Axioms $\mathbf{4}_{\mathtt{Bel}}$ and $\mathbf{5}_{\mathtt{Bel}}$), and we assume that an agent cannot have inconsistent beliefs and conflicting choices (Axioms $\mathbf{D}_{\mathtt{Bel}}$ and $\mathbf{D}_{\mathtt{Choice}}$). Axiom $\mathbf{K}_{\mathtt{After}}$ says that every modal operator $\mathtt{After}_{i:\alpha}$ obeys the principles of the basic normal modal logic K, Axiom $\mathbf{K}_{\mathtt{Does}}$ says the same for every $\mathtt{Does}_{i:\alpha}$, and Axiom $\mathbf{K}_{\mathtt{G}}$ says the same for the temporal operator $\mathtt{G}$. Axiom $\mathbf{Alt}_{\mathtt{Does}}$ says that if i is going to do α and φ will be true afterward, then it cannot be the case that j is going to do β and $\neg\varphi$ will be true afterward. The axioms for time $\mathbf{4}_{\mathtt{G}}$ and $\mathbf{H}_{\mathtt{G}}$ correspond to the transitivity and connectedness constraint of the relation G. The other axioms are about more complex interactions between the modal operators, and will be discussed in detail in the rest of the section.

3.4 Discussion of interaction axioms

In the rest of the section we discuss the relevant interactions between modal operators of our logic.

3.4.1 Actions and intentions

Axiom $\mathbf{Inc}_{\mathtt{After},\mathtt{Does}}$ says that if i is going to do α and φ will be true afterward, then it is not the case that φ will be false after i does α. Axioms **IntAct1** and **IntAct2** relate goals with actions. According to **IntAct1**, if i has the goal to do action α and has the capacity to do α, then i is going to do α. According to **IntAct2**, an agent is going to do action α only if it has the goal to do α. We therefore suppose that an agent's *doing* is by definition intentional.

(**PC**)	all theorems of propositional calculus
(**K**$_{\mathtt{Bel}}$)	$(\mathtt{Bel}_i\varphi \wedge \mathtt{Bel}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Bel}_i\psi$
(**D**$_{\mathtt{Bel}}$)	$\neg(\mathtt{Bel}_i\varphi \wedge \mathtt{Bel}_i\neg\varphi)$
(**4**$_{\mathtt{Bel}}$)	$\mathtt{Bel}_i\varphi \rightarrow \mathtt{Bel}_i\mathtt{Bel}_i\varphi$
(**5**$_{\mathtt{Bel}}$)	$\neg\mathtt{Bel}_i\varphi \rightarrow \mathtt{Bel}_i\neg\mathtt{Bel}_i\varphi$
(**K**$_{\mathtt{Choice}}$)	$(\mathtt{Choice}_i\varphi \wedge \mathtt{Choice}_i(\varphi \rightarrow \psi)) \rightarrow \mathtt{Choice}_i\psi$
(**D**$_{\mathtt{Choice}}$)	$\neg(\mathtt{Choice}_i\varphi \wedge \mathtt{Choice}_i\neg\varphi)$
(**4**$_{\mathtt{Choice}}$)	$\mathtt{Choice}_i\varphi \rightarrow \mathtt{Bel}_i\mathtt{Choice}_i\varphi$
(**5**$_{\mathtt{Choice}}$)	$\neg\mathtt{Choice}_i\varphi \rightarrow \mathtt{Bel}_i\neg\mathtt{Choice}_i\varphi$
(**K**$_{\mathtt{After}}$)	$(\mathtt{After}_{i:\alpha}\varphi \wedge \mathtt{After}_{i:\alpha}(\varphi \rightarrow \psi)) \rightarrow \mathtt{After}_{i:\alpha}\psi$
(**K**$_{\mathtt{Does}}$)	$(\mathtt{Does}_{i:\alpha}\varphi \wedge \neg\mathtt{Does}_{i:\alpha}\neg\psi) \rightarrow \mathtt{Does}_{i:\alpha}(\varphi \wedge \psi)$
(**Alt**$_{\mathtt{Does}}$)	$\mathtt{Does}_{i:\alpha}\varphi \rightarrow \neg\mathtt{Does}_{j:\beta}\neg\varphi$
(**K**$_{\mathtt{G}}$)	$(\mathtt{G}\varphi \wedge \mathtt{G}(\varphi \rightarrow \psi)) \rightarrow \mathtt{G}\psi$
(**4**$_{\mathtt{G}}$)	$\mathtt{G}\varphi \rightarrow \mathtt{GG}\varphi$
(**H**$_{\mathtt{G}}$)	$(\mathtt{F}\varphi \wedge \mathtt{F}\psi) \rightarrow (\mathtt{F}(\varphi \wedge \mathtt{F}\psi) \vee \mathtt{F}(\psi \wedge \mathtt{F}\varphi) \vee \mathtt{F}(\psi \wedge \varphi))$
(**Inc**$_{\mathtt{After},\mathtt{Does}}$)	$\mathtt{Does}_{i:\alpha}\varphi \rightarrow \neg\mathtt{After}_{i:\alpha}\neg\varphi$
(**IntAct1**)	$(\mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top \wedge \neg\mathtt{After}_{i:\alpha}\bot) \rightarrow \mathtt{Does}_{i:\alpha}\top$
(**IntAct2**)	$\mathtt{Does}_{i:\alpha}\top \rightarrow \mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top$
(**WR**)	$\mathtt{Bel}_i\varphi \rightarrow \neg\mathtt{Choice}_i\neg\varphi$
(**Inc**$_{\mathtt{G},\mathtt{Does}}$)	$\mathtt{Does}_{i:\alpha}\varphi \rightarrow \mathtt{F}\varphi$
(**OneStepAct**)	$\mathtt{Does}_{i:\alpha}\mathtt{G}^*\varphi \rightarrow \mathtt{G}\varphi$
(**MP**)	from φ and $\varphi \rightarrow \psi$ infer ψ
(**Nec**$_{\mathtt{Bel}}$)	from φ infer $\mathtt{Bel}_i\varphi$
(**Nec**$_{\mathtt{Choice}}$)	from φ infer $\mathtt{Choice}_i\varphi$
(**Nec**$_{\mathtt{G}}$)	from φ infer $\mathtt{G}\varphi$
(**Nec**$_{\mathtt{After}}$)	from φ infer $\mathtt{After}_{i:\alpha}\varphi$
(**Nec**$_{\mathtt{Does}}$)	from φ infer $\neg\mathtt{Does}_{i:\alpha}\neg\varphi$

Figure 1: Axiomatization of $\mathcal{L}$

Given $\mathbf{Inc}_{\mathtt{After},\mathtt{Does}}$, **IntAct1** and **IntAct2** could be replaced by the single axiom:

(**IntAct**) $\qquad \mathtt{Does}_{i:\alpha}\top \leftrightarrow (\mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top \wedge \neg\mathtt{After}_{i:\alpha}\bot)$

Similar axioms have been studied in [48] in which a logical model of the relationships between intention and action performance is proposed.

3.4.2 Time and actions

Axioms $\mathbf{Inc}_{\mathtt{G},\mathtt{Does}}$ and **OneStepAct** describe the interaction between actions and time. According to Axiom $\mathbf{Inc}_{\mathtt{G},\mathtt{Does}}$, the outcome that results from the execution of α by i is in the future of the current state. According to Axiom **OneStepAct**, if agent i is going to do action α, and after i's action φ will be immediately true and will be always true then, φ will be always true. This axiomatizes that every action occurrence takes one time step.

3.4.3 Beliefs and choices

As far as beliefs and chosen goals (choices) are concerned, we suppose that the two kinds of mental attitudes must be compatible, that is, if an agent has the goal that φ, then it cannot believe that $\neg\varphi$. This is the so-called assumption of *weak realism* [8, 53]. According to this hypothesis, a rational agent cannot choose φ if it believes that φ is an impossible state of affairs.[6] The principle of weak realism is expressed by Axiom **WR**.

We also assume positive and negative introspection over chosen goals, as expressed by axioms $\mathbf{4}_{\mathtt{Choice}}$ and $\mathbf{5}_{\mathtt{Choice}}$. Together with $\mathbf{D}_{\mathtt{Choice}}$ they imply the equivalences $\mathtt{Choice}_i\varphi \leftrightarrow \mathtt{Bel}_i\mathtt{Choice}_i\varphi$ and $\neg\mathtt{Choice}_i\varphi \leftrightarrow \mathtt{Bel}_i\neg\mathtt{Choice}_i\varphi$.

3.5 Soundness and completeness

We call $\mathcal{L}$ the logic axiomatized by the axioms and rules of inference presented above. We write $\vdash_{\mathcal{L}} \varphi$ if formula φ is a theorem of $\mathcal{L}$ (i.e. φ is derivable from the axioms and rules of inference of the logic $\mathcal{L}$). An example of $\mathcal{L}$ theorem, which follows from **IntAct1**, **IntAct2** and $\mathbf{Inc}_{\mathtt{After},\mathtt{Does}}$, is the following:

(**IntAct**) $\qquad \mathtt{Does}_{i:\alpha}\top \leftrightarrow (\mathtt{Choice}_i\mathtt{Does}_{i:\alpha}\top \wedge \neg\mathtt{After}_{i:\alpha}\bot)$

that we have already introduced in Section 3.4.1.

According to Theorem **IntAct**, an agent i does an action α if and only if, i is capable to do α and intends to do it. This highlights that in our framework actions are supposed to be intentional.

We can prove that the logic $\mathcal{L}$ is *sound* and *complete* with respect to the class of $\mathcal{L}$-frames. Namely:

Theorem 1. *$\mathcal{L}$ is determined by the class of $\mathcal{L}$-frames.*

[6] An alternative hypothesis proposed in [2, 29] is the so-called *strong realism* principle. According to this principle, an agent cannot choose φ unless it believes that φ will occur. Strong realism has been widely criticized (e.g. in [8]) and we prefer not to consider it here.

Proof. It is a routine task to check that the axioms of the logic $\mathcal{L}$ correspond one-to-one to their semantic counterparts on the frames. In particular, Axioms $\mathbf{D}_{\mathtt{Bel}}$, $\mathbf{4}_{\mathtt{Bel}}$, $\mathbf{5}_{\mathtt{Bel}}$ correspond (in the sense of correspondence theory, see for instance [64, 6]) to the seriality, transitivity and Euclideanity of every relation B_i. Axiom $\mathbf{D}_{\mathtt{Choice}}$ corresponds to the seriality of every relation C_i. Axiom $\mathbf{4}_{\mathtt{G}}$ corresponds to the transitivity of the relation G, and Axiom $\mathbf{H}_{\mathtt{G}}$ to connectedness. Axiom $\mathbf{Alt}_{\mathtt{Does}}$ corresponds to the semantic constraint **S1**. Axiom $\mathbf{Inc}_{\mathtt{After},\mathtt{Does}}$ corresponds to the semantic constraint **S2**. Axioms **IntAct1** and **IntAct2**, correspond to the constraints **S3** and **S4**. Axioms $\mathbf{Inc}_{\mathtt{G},\mathtt{Does}}$ and **OneStepAct** correspond to the constraints **S5** and **S6**. Axiom **WR** corresponds to the constraint **S7**. Finally, Axioms $\mathbf{4}_{\mathtt{Choice}}$ and $\mathbf{5}_{\mathtt{Choice}}$ correspond together to the constraint **S8**.

It is routine, too, to check that all of our axioms are in the Sahlqvist class. This means that the axioms are all expressible as first-order conditions on frames and that they are complete with respect to the defined frames classes, cf. [6, Th. 2.42]. □

4 Trust: two formalizations

In the next two subsections we are going to formally define occurrent and dispositional trust by means of two formula schemas of our logic $\mathcal{L}$.

4.1 Occurrent trust

Following C&F theory as exposed in Section 2, we define occurrent trust $\mathtt{OccTrust}(i,j,\alpha,\varphi)$ as the conjunction of $\mathtt{Choice}_i\mathtt{F}\varphi$ and $\mathtt{Bel}_i(\mathtt{Intends}_j(\alpha)\wedge\mathtt{Capable}_j(\alpha)\wedge\mathtt{After}_{j:\alpha}\varphi)$. Note that the notion of goal in the definition of occurrent trust, i.e. $\mathtt{Choice}_i\mathtt{F}\varphi$, has to be about the strict future, and should not include the present: trust is about the effect φ of j's action that only obtains at the next time point, and $Choice_i\ \varphi$ should not be relevant here. To see this, consider a situation where i has goal that φ now, but has goal that φ be impossible in the strict future ($\mathtt{Choice}_i\varphi\wedge\mathtt{Choice}_i\neg\mathtt{F}\varphi$): in such a situation there is no point in i trusting j to achieve φ!

As we have seen, in our logic the principle of intentional action

$$\mathtt{Does}_{j:\alpha}\top\leftrightarrow\mathtt{Capable}_j(\alpha)\wedge\mathtt{Intends}_j(\alpha)$$

holds (Theorem **IntAct**). The definition of occurrent trust can therefore be simplified to:

$$\mathtt{OccTrust}(i,j,\alpha,\varphi)\stackrel{\text{def}}{=}\mathtt{Choice}_i\mathtt{F}\varphi\wedge\mathtt{Bel}_i(\mathtt{Does}_{j:\alpha}\top\wedge\mathtt{After}_{j:\alpha}\varphi)$$

This formalization better expresses the fundamental aspect of the concept of occurrent trust, namely the fact that i trusts j to do α with respect to φ if and only if, i wants φ to be true at some point in the future and believes that the trustee will ensure φ by doing action α.

It might be argued that it is too strong to require that j is going to do α immediately, and that it is sufficient to simply require that j will do α eventually, i.e. $\mathtt{Bel}_i\mathtt{F}^*(\mathtt{Does}_{j:\alpha}\top\wedge\mathtt{After}_{j::\alpha}\varphi)$. However, '$j$ acting here and now' is part of C&F's original definition, and —at least as far as occurrent trust is concerned— the aim of the paper is to provide a formalization of their concept. Moreover, note that if we just

require α to be done at some point in the future then we get a problem with procrastination: i should not trust j if he believes that j is systematically postponing his action. If we want to exclude this then we have to integrate deadlines (which we wanted to avoid in order to keep things simple).

EXAMPLE. Suppose that Bill trusts Mary to send him a certain product with regard to his goal of possessing the product:

$$\texttt{OccTrust}(Bill, Mary, send_{Bill}, hasProduct_{Bill}).$$

This means that Bill wants to receive the product at some point in the future:

$$\texttt{Choice}_{Bill}\texttt{F}hasProduct_{Bill}.$$

Moreover, according to Bill's beliefs, Mary, by sending the product, will ensure that he will receive the product, and Mary is going to send the product to Bill:

$$\texttt{Bel}_{Bill}(\texttt{Does}_{Mary:send_{Bill}}\top \wedge \texttt{After}_{Mary:send_{Bill}} hasProduct_{Bill}).$$

The following theorems highlight some interesting properties of the notion of occurrent trust.

Theorem 2. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(2a) $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi$

(2b) $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{F}\varphi$

(2c) $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \leftrightarrow \texttt{Bel}_i\texttt{OccTrust}(i,j,\alpha,\varphi)$

(2d) $\vdash_{\mathcal{L}} (\texttt{OccTrust}(i,j,\alpha,\varphi) \vee \texttt{OccTrust}(i,j,\alpha,\psi)) \rightarrow \texttt{OccTrust}(i,j,\alpha,\varphi \vee \psi)$

(2e) $\vdash_{\mathcal{L}} \texttt{Bel}_i\neg\texttt{Capable}_j(\alpha) \rightarrow \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$

(2f) $\vdash_{\mathcal{L}} \texttt{Bel}_i\neg\texttt{Intends}_j(\alpha) \rightarrow \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$

(2g) $\vdash_{\mathcal{L}} \texttt{Bel}_i\neg\texttt{After}_{j:\alpha}\varphi \rightarrow \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$

(2h) $\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \rightarrow \neg\texttt{OccTrust}(i,j,\alpha,\varphi)$

(2i) $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\top) \leftrightarrow (\texttt{Bel}_i\texttt{Intends}_j(\alpha) \wedge \texttt{Choice}_i\texttt{F}\top)$

Proof. All the proofs are straightforward.
2a. First, $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top)$.
Second, $\vdash_{\mathcal{L}} \texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top \rightarrow \texttt{Does}_{j:\alpha}\varphi$ by Axiom $\mathbf{Inc}_{\texttt{After},\texttt{Does}}$ and standard principles for the normal operator $\texttt{Does}_{j:\alpha}$.
2b. First, Theorem 2a guarantees that $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi$.
Second, $\vdash_{\mathcal{L}} \texttt{Does}_{j:\alpha}\varphi \rightarrow \texttt{F}\varphi$ by Axiom $\mathbf{Inc}_{\texttt{G},\texttt{Does}}$.
We conclude by Axiom $\mathbf{K}_{\texttt{Bel}}$ that $\vdash_{\mathcal{L}} \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top) \rightarrow \texttt{Bel}_i\texttt{F}\varphi$.
2c. We have $\vdash_{\mathcal{L}} \texttt{Choice}_i\texttt{F}\varphi \leftrightarrow \texttt{Bel}_i\texttt{Choice}_i\texttt{F}\varphi$ by $\mathbf{D}_{\texttt{Choice}}$, $\mathbf{4}_{\texttt{Choice}}$ and $\mathbf{5}_{\texttt{Choice}}$, and we have
$\vdash_{\mathcal{L}} \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top) \leftrightarrow \texttt{Bel}_i\texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top)$
by $\mathbf{D}_{\texttt{Bel}}$, $\mathbf{4}_{\texttt{Bel}}$ and $\mathbf{5}_{\texttt{Bel}}$.
2d. We prove that $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{OccTrust}(i,j,\alpha,\varphi \vee \psi)$. To see that it suffices to observe that $\vdash_{\mathcal{L}} \texttt{Choice}_i\texttt{F}\varphi \rightarrow \texttt{Choice}_i\texttt{F}(\varphi \vee \psi)$ and that
$\vdash_{\mathcal{L}} \texttt{Bel}_i(\texttt{After}_{j:\alpha}\varphi \wedge \texttt{Does}_{j:\alpha}\top) \rightarrow \texttt{Bel}_i(\texttt{After}_{j:\alpha}(\varphi \vee \psi) \wedge \texttt{Does}_{j:\alpha}\top)$;
both actually follow from standard principles of normal modal logics.

2e. First, $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{Capable}_j(\alpha)$ by definition of occurrent trust (Theorem **IntAct**).
Second, $\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{Capable}_j(\alpha) \rightarrow \neg\texttt{Bel}_i\neg\texttt{Capable}_j(\alpha)$ by Axiom $\mathbf{D}_{\texttt{Bel}}$.
It follows that $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \neg\texttt{Bel}_i\neg\texttt{Capable}_j(\alpha)$, from which the theorem follows by contraposition.
2f. The proof is similar to that of Theorem 2e, replacing $\texttt{Capable}_j(\alpha)$ by $\texttt{Intends}_j(\alpha)$.
2g. We can prove $\vdash_{\mathcal{L}} \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{After}_{j:\alpha}\varphi$ in a way similar to the two preceding proofs.
2h. Theorem 2a together with standard modal principles tells us that
$\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi$.
Then
$\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi \rightarrow \texttt{Bel}_i\texttt{After}_{j:\alpha}\bot$
by $\mathbf{Inc}_{\texttt{After},\texttt{Does}}$ and standard modal principles for $\texttt{After}_{j:\alpha}$. Then by $\mathbf{Alt}_{\texttt{Does}}$ and $\mathbf{D}_{\texttt{Bel}}$ it follows that
$\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\neg\varphi \wedge \texttt{Bel}_i\texttt{Does}_{j:\alpha}\varphi \rightarrow \bot$.
2i. We use that due to necessitation for $\texttt{After}_{j:\alpha}$ and $\texttt{Bel}_i$ ($\mathbf{Nec}_{\texttt{Does}}$ and $\mathbf{Nec}_{\texttt{Bel}}$) we have $\vdash_{\mathcal{L}} \texttt{After}_{j:\alpha}\top$ and $\vdash_{\mathcal{L}} \texttt{Bel}_i\texttt{After}_{j:\alpha}\top$.

□

According to Theorem 2b, if i trusts j to do α with regard to φ then i has a positive expectation that φ will be true at some point in the future. Theorem 2c highlights the fact that trust is under the focus of the truster's awareness: i trusts j to do α with regard to φ if and only if i is aware of this. According to Theorem 2d if i trusts j to do α with regard to φ or i trusts j to do α with regard to ψ then, i trusts j to do α with regard to φ or ψ. The other direction is not valid. For instance, i might trust j to send the product i has payed so that it will receive the product either tomorrow or the day after tomorrow. This does not necessarily imply that either i trusts j to send the product so that it will receive the product tomorrow, or i trusts j to send the product so that it will receive the product the day after tomorrow (i might be uncertain about the day it will receive the product). The three theorems 2e-2g shows that agent i cannot trust agent j to do α with respect to φ if one of the three relevant properties of the trustee is lacking (j's capability to do α, j's intention to do α or j's opportunity to ensure φ by doing α). Finally, according to Theorem 2h, i cannot trust agent j to do α with respect to φ if i believes that j has the opportunity to ensure $\neg\varphi$ by doing α.

Trust in the trustee's action vs. trust in the trustee's inaction Elsewhere [47] we have distinguished *trust in the trustee's action* from *trust in the trustee's inaction*. While the former concept is focused on the domain of gains, the latter is focused on the domain of losses. That is, in the former case the truster believes that the trustee is in condition to *further* the achievement (or the maintenance) of a pleasant state of affairs, and it will *do* that; in the latter case the truster believes that the trustee is in condition to *endanger* the achievement (or the maintenance) of a pleasant state of affairs, but it will *refrain* from doing that.[7] According to the proposed definition, i trusts j *not to do*

[7] This opposition is symmetrical to the opposition between *doing* and *refraining* (or *forbearing*) which has been studied in the philosophy of action [65, 5].

α with regard to φ if and only if i wants φ to be true at some point in the future and i believes that j, by doing α, will ensure φ to be always false, j has the capacity to do α, but j does not intend to do α. That is,

$$\texttt{OccTrust}(i,j,\sim\alpha,\varphi) \stackrel{\text{def}}{=}$$

$$\texttt{Choice}_i\texttt{F}\varphi \wedge \texttt{Bel}_i(\texttt{Capable}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\texttt{G}^*\neg\varphi \wedge \neg\texttt{Intends}_j(\alpha))$$

Differently from trust in the trustee's action, trust in the trustee's inaction does not necessarily imply a positive expectation of the truster. More precisely, i might trust j not to do α with regard to φ, without believing that φ will be true in the next state, i.e. $\texttt{OccTrust}(i,j,\sim\alpha,\varphi) \wedge \neg\texttt{Bel}_i\texttt{F}\varphi$ is satisfiable in our logic. The intuitive reason is that i can trust j not to do α with regard to φ and, at the same time, believe that another agent will prevent φ from being true at some point in the future.

It is to be noted that we could make our trust model more sophisticated by distinguishing two kinds of trust in the trustee's action and two kinds of trust in the trustee's inaction depending on the kind of goal that is involved. A classical distinction in BDI approaches is the distinction between achievement goal and maintenance goal. An agent i has an achievement goal that φ if and only if, agent i does not believe that φ holds now and has the goal to achieve φ at some point in the future. An agent i has a maintenance goal that φ if and only if i believes that φ already holds, and has the goal to continue to have φ in the future. Formally,

$$\texttt{AGoal}_i\varphi \stackrel{\text{def}}{=} \texttt{Choice}_i\texttt{F}\varphi \wedge \neg\texttt{Bel}_i\varphi$$
$$\texttt{MGoal}_i\varphi \stackrel{\text{def}}{=} \texttt{Choice}_i\texttt{G}\varphi \wedge \texttt{Bel}_i\varphi.$$

Given this distinction between achievement goal and maintenance goal, four cases of occurrent trust can be identified just by substituting $\texttt{Choice}_i\texttt{F}\varphi$ either with $\texttt{AGoal}_i\varphi$ or with $\texttt{MGoal}_i\varphi$ in the definition of $\texttt{OccTrust}(i,j,\alpha,\varphi)$ and of $\texttt{OccTrust}(i,j,\sim\alpha,\varphi)$.

1. i trusts j to do α with respect to the achievement of φ,
2. i trusts j to do α with respect to the maintenance of φ,
3. i trusts j not to do α with respect to the achievement of φ,
4. i trusts j not to do α with respect to the maintenance of φ.

An example of the first case is the occurrent trust that an agent i has in an agent j, when i wants to receive a product from j that it has not yet received, and believes that j is going to send it the product so that it will receive it. An example of the second case is the occurrent trust that an important politician might have in her bodyguard: the politician is currently alive and believes that his bodyguard will contribute to the maintenance of this state of affairs by defending him against a fatal terroristic attack. An example of the third case is the occurrent trust that an agent i might have in another agent j during a trade transaction. Imagine i wants to buy a certain product from j at a price lower than 500 Euros, and i is trying to convince j. In this situation, agent i believes that j is capable to refuse i's offer and, by refusing i's offer, j will prevent i from purchasing the product at a price lower than 500 Euros. But i believes that j does not intend to refuse its offer. An example of the fourth case is the occurrent trust

that during Cold War the president of the USA had in that of the USSR: USA believe that the Earth has not been destroyed yet, but USSR have the opportunity to destroy the Earth forever by launching an atomic attack against USA. Moreover USA believe that USSR are capable to launch an atomic attack against USA, but do not intend to do that.

4.2 Dispositional trust

As emphasized in Section 2, dispositional trust is trust in a general disposition of the trustee to perform an instance of the action type α. More precisely, agent i is disposed to trust agent j if and only if, i thinks that it will possibly need j's action α in the future to solve a certain task φ, and whenever it will need j's action α to achieve φ and j will be required to perform α, j will perform α so that φ will be ensured.

Dispositional trust is defined with respect to a weaker concept of goal than the one involved in occurrent trust. We call it potential goal. A potential goal φ in context κ is a goal which the agent does not exclude to have one day as a goal:

$$\texttt{PotGoal}_i(\varphi,\kappa) \stackrel{\text{def}}{=} \texttt{Poss}_i\texttt{F}^*(\kappa \wedge \texttt{Choice}_i\texttt{F}\varphi)$$

More precisely, i has the potential goal that φ in given circumstances κ if and only if, according to i, it is possible that at some point in the future in which κ holds it will want φ to be true. For example, while being currently in good health, Bill might have the potential goal of recovering, i.e. $\texttt{PotGoal}_i(recovered, asked_{Bill,treat,doc})$.

Dispositional trust has five arguments: a truster i, a trustee j, an action α of j, a potential goal φ of i, and certain circumstances κ.

We say that "i is disposed to trust j to do α with respect to φ in the circumstances κ", if and only if:

1. i has the potential goal φ in given circumstances κ;
2. i believes that always, if it wants φ to be true and κ holds, then
   (a) j will be capable to do α;
   (b) j, by doing α, will ensure φ; and
   (c) j will intend to do α.

The element κ in the definition of dispositional trust is used to describe the circumstances under which the truster expects that the trustee will ensure φ by doing α. For example, in the context of a trade transaction between agent i and agent j, κ denotes that agent i has paid a certain product to j and has a proof for that. That is, i is disposed to trust j to send a certain product under the condition that i has paid the product to j and that i has a proof for that.

When dispositional trust is based on norms, the condition κ expresses the general fact "agent j is required to perform action α for i". This is just a different way to say that j has a *directed obligation* towards i to do action α (see e.g. [41, 45, 50] for a logical account of directed obligations). By way of example, imagine the situation in which i is an employer of a certain company and j is an employee. Agent i's trust

in j with respect to the accomplishment of a given task φ is based on i's belief that, whenever j is required to accomplish the task and i wants the task to be accomplished, j will do as required.

In formulas, dispositional trust $\texttt{DispTrust}(i,j,\alpha,\varphi,\kappa)$ is therefore defined as the conjunction of $\texttt{PotGoal}_i(\varphi,\kappa)$ and

$$\texttt{Bel}_i\texttt{G}^*((\kappa \wedge \texttt{Choice}_i\texttt{F}\varphi) \rightarrow (\texttt{Intends}_j(\alpha) \wedge \texttt{Capable}_j(\alpha) \wedge \texttt{After}_{j:\alpha}\varphi)).$$

As for occurrent trust, the conditions $\texttt{Capable}_j(\alpha)$ and $\texttt{Intends}_j(\alpha)$ together are equivalent to $\texttt{Does}_{j:\alpha}\top$ (Theorem **IntAct**), and the definition of dispositional trust can be simplified to:

$$\texttt{DispTrust}(i,j,\alpha,\varphi,\kappa) \stackrel{\text{def}}{=} \texttt{PotGoal}_i(\varphi,\kappa) \wedge$$

$$\texttt{Bel}_i\texttt{G}^*((\kappa \wedge \texttt{Choice}_i\texttt{F}\varphi) \rightarrow (\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi))$$

EXAMPLE. Suppose that Bill is disposed to trust his doctor to treat him in order to cure him of a certain illness, in the circumstances in which he asks the doctor to treat him:

$$\texttt{DispTrust}(Bill, doc, treat, recovered, asked_{Bill,treat,doc}).$$

This means that Bill has the potential goal of recovering from some illness, in the circumstances in which he asks the doctor to treat him:

$$\texttt{PotGoal}_{Bill}(recovered, asked_{Bill,treat,doc}).$$

Moreover, Bill believes that whenever he will want to recover from some illness and will ask the doctor to treat him, the doctor will do the treatment so that he will recover:

$$\texttt{Bel}_{Bill}\texttt{G}^*((\texttt{Choice}_{Bill}\texttt{F}recovered \wedge asked_{Bill,treat,doc}) \rightarrow (\texttt{Does}_{doc:treat}\top \wedge \texttt{After}_{doc:treat}recovered)).$$

The following theorem highlights the relationships between dispositional trust and occurrent trust.

Theorem 3. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(3) $\quad \vdash_{\mathcal{L}} (\texttt{DispTrust}(i,j,\alpha,\varphi,\kappa) \wedge \texttt{Choice}_i\texttt{F}\varphi \wedge \texttt{Bel}_i\kappa) \rightarrow \texttt{OccTrust}(i,j,\alpha,\varphi)$

Proof. $\texttt{DispTrust}(i,j,\alpha,\varphi,\kappa) \wedge \texttt{Choice}_i\texttt{F}\varphi \wedge \texttt{Bel}_i\kappa$ implies

$$\texttt{Choice}_i\texttt{F}\varphi \wedge \texttt{Bel}_i(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa) \wedge$$
$$\texttt{Bel}_i\texttt{G}^*((\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa) \rightarrow (\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi))$$

(by Axiom $\mathbf{4}_{\texttt{Choice}}$ and standard principles of the normal operator $\texttt{Bel}_i$). From the latter it follows that $\texttt{Choice}_i\texttt{F}\varphi \wedge \texttt{Bel}_i(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi)$ (by Axioms $\mathbf{K}_{\texttt{Bel}}$ and definition of $\texttt{G}^*$) which in turn implies $\texttt{OccTrust}(i,j,\alpha,\varphi)$. □

According to Theorem 3, if i is disposed to trust j with respect to φ in the circumstances κ, i wants φ to be true and believes that κ holds then, i trusts j to do α with respect to φ.

The following theorems highlight other interesting properties of the notion of dispositional trust.

Theorem 4. *Let $i, j \in AGT$ and $\alpha \in ACT$. Then:*

(4a) $\quad \vdash_{\mathcal{L}} (\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa) \wedge \mathtt{Choice}_i\mathsf{F}\varphi \wedge \mathtt{Bel}_i\kappa) \rightarrow \mathtt{Bel}_i\mathsf{F}\varphi$

(4b) $\quad \vdash_{\mathcal{L}} \mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa) \leftrightarrow \mathtt{Bel}_i\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa)$

Proof. To prove Theorem 4a, first observe that by Theorem 3 we have $\vdash_{\mathcal{L}} (\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa) \wedge \mathtt{Choice}_i\mathsf{F}\varphi \wedge \mathtt{Bel}_i\kappa) \rightarrow \mathtt{OccTrust}(i,j,\alpha,\varphi)$. Then by Theorem 2a we have $\vdash_{\mathcal{L}} \mathtt{OccTrust}(i,j,\alpha,\varphi) \rightarrow \mathtt{Bel}_i\mathtt{Does}_{j:\alpha}\varphi$, and the result follows by propositional principles.

Now let us prove Theorem 4b. By definition, $\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa)$ is: $\mathtt{Poss}_i\mathsf{F}(\kappa \wedge \mathtt{Choice}_i\mathsf{F}\varphi) \wedge \mathtt{Bel}_i\mathsf{G}^*((\kappa \wedge \mathtt{Choice}_i\mathsf{F}\varphi) \rightarrow (\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{After}_{j:\alpha}\varphi))$. Since $\mathtt{Bel}_i\varphi \leftrightarrow \mathtt{Bel}_i\mathtt{Bel}_i\varphi$ and $\mathtt{Poss}_i\varphi \leftrightarrow \mathtt{Bel}_i\mathtt{Poss}_i\varphi$ are $\mathcal{L}$ theorems (by Axioms $\mathbf{D}_{\mathtt{Bel}}$, $\mathbf{4}_{\mathtt{Bel}}$ and $\mathbf{5}_{\mathtt{Bel}}$), we conclude that the latter is equivalent to $\mathtt{Bel}_i\mathtt{Poss}_i\mathsf{F}(\kappa \wedge \mathtt{Choice}_i\mathsf{F}\varphi) \wedge \mathtt{Bel}_i\mathtt{Bel}_i\mathsf{G}^*((\mathtt{Choice}_i\mathsf{F}\varphi \wedge \kappa) \rightarrow (\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{After}_{j:\alpha}\varphi))$. By standard principles of the modal operator $\mathtt{Bel}_i$, the latter formula is equivalent to $\mathtt{Bel}_i\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa)$. □

According to Theorem 4a, if i is disposed to trust j with respect to φ in the circumstances κ, i wants φ to be true at some point in the future and believes that κ is true, then i has a positive expectation that φ will be true at some point in the future. Theorem 4b highlights that an agent is correctly aware of its disposition to trust someone. Similar properties have been discussed for occurrent trust (see Theorem 2). It is straightforward to prove that dispositional trust does not aggregate under conjunction, i.e. the formula

$(\mathtt{DispTrust}(i,j,\alpha,\varphi,\kappa) \wedge \mathtt{DispTrust}(i,j,\alpha,\psi,\kappa)) \rightarrow \mathtt{DispTrust}(i,j,\alpha,\varphi \wedge \psi,\kappa)$

is invalid in our logic. This is due to the definition of potential goal: i might be disposed to trust j with respect to φ and might be disposed to trust j with respect to ψ, and expect that it will have the goal φ and the goal ψ at different moments in the future. Thus, i is not disposed to trust j with respect to $\varphi \wedge \psi$.

REMARK. Note that as for occurrent trust we might define dispositional trust in the trustee's inaction. We say that "i is disposed to trust j not to do α with respect to φ in the circumstances κ", if and only if i has the potential goal φ in given circumstances κ, i believes that always, if it wants φ to be true in the future and j is capable to do α and j has the opportunity to ensure φ to be always false by doing α, then j does not intend to do α. Formally,

$$\mathtt{DispTrust}(i,j,\sim\alpha,\varphi,\kappa) \stackrel{\text{def}}{=} \mathtt{PotGoal}_i(\varphi,\kappa) \wedge$$

$$\mathtt{Bel}_i\mathsf{G}^*((\mathtt{Choice}_i\mathsf{F}\varphi \wedge \mathtt{Capable}_j(\alpha) \wedge \mathtt{After}_{j:\alpha}\mathsf{G}^*\neg\varphi \wedge \kappa) \rightarrow \neg\mathtt{Intends}_j(\alpha))$$

4.3 Discussion

Many different views and definitions of trust have been proposed in several disciplines such as economics, philosophy, sociology, computer science, and many types of trust

that have been considered, that are more or less general. Here we focus on C&F's trust definition, refining their approach by distinguishing occurrent from dispositional trust. We think that C&F's definition is one of the most general and is very well-suited to be used in a BDI-like framework in which agents are defined in terms of their mental attitudes (such as beliefs, goals and intentions). We here compare the trust definitions presented in Sections 4.1 and 4.2 and inspired by C&F with some alternative definitions studied in the literature on trust and reputation.

One of the most popular and general definitions of trust is that of Deutsch [20, 21]. A further elaboration of this definition is given in [26] where trust is conceived as "[...] a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action." [26, p. 217] In Deutsch's perspective, trust involves risk perception, that is, when i trusts j, i thinks it to be possible that j will perform a certain action α whose occurrence will lead to positive motivational consequences (i.e., achievement of some goal) and, at the same time, i thinks it to be possible that j will not perform that action, so that there will be negative motivational consequences (the frustration of some goal). In this sense, when i trusts j, i is doing a real bet on j. In the definitions of occurrent trust and dispositional trust presented in Sections 4.1 and 4.2, such a dimension of risk perception is not taken into account. Note that this dimension could be captured by extending our logical framework with operators of graded belief of the form $\mathtt{Bel}_i^x$ as the ones studied in [30, 22]. The formula $\mathtt{Bel}_i^x\varphi$ expresses that agent i thinks φ to be probable with degree x, with $x \in [0,1]$. In this extended logical framework, we could express that i (the truster) thinks it to be probable with degree x that j (the trustee) will perform action α, and thinks it to be probable with degree $1-x$ that the trustee will not perform action α. Moreover, i believes that j, by doing α, will ensure φ and believes that if j will not do α, φ will not be achieved. That is, $\mathtt{Bel}_i^x\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{Bel}_i^{1-x}\neg\mathtt{Does}_{j:\alpha}\top \wedge \mathtt{Bel}_i\mathtt{After}_{j:\alpha}\varphi \wedge \mathtt{Bel}_i(\neg\mathtt{Does}_{j:\alpha}\top \rightarrow \mathtt{G}\neg\varphi)$ where $\mathtt{Bel}_i\varphi$ is identified with $\mathtt{Bel}_i^1\varphi$. In this situation, while trusting j to do action α with respect to φ, i perceives the risk of failure, that is, it thinks it to be probable with degree $1-x$ that j will not perform α so that the goal φ will be frustrated.

Some authors [3, 34] have focused on a very specific notion of trust (called "genuine trust") based on goal adoption and on the truster's belief that the trustee has goodwill towards it (see also [10] for a discussion of this notion trust). In this perspective, an agent i is said to trust agent j only if i has a certain goal φ and believes that j will adopt its goal and will decide to promote it, so that i's goal will prevail (in case of conflict) over the other goals of j. That is, i trusts j only if, i wants a state of affairs φ to be true and believes that j will act in order to promote φ *because* j believes that i wants φ to be true and *because* j wants that i will achieve φ (i.e. j's *reason* for promoting φ is j's belief that i wants φ to be true and j's goal that i will achieve φ). This form of "genuine trust" based on goal adoption and on the trustee's goodwill towards the truster is similar to the notion of trust recently studied in the context of game theory (see e.g. [4, 55]). One of the basic features of the so-called Trust Game, as isolated in [55], is the temptation for j (the trustee) to be opportunistic, after i (the truster) has decided to rely on j. In this kind of game, after i has decided to rely on j, j can choose either an

action α that leads to very good consequences for j and to very bad consequences for i, or an action β that leads to worse consequences for j and to better consequences for i. Agent i is said to trust j only if, i believes that j will do β, that is, i believes that j will not be opportunistic, but it will rather promote i's interests. This notion of "genuine trust" is obviously stronger than our notions of occurrent and dispositional trust. Consider for instance occurrent trust. In our definition we do not require that, according to i (the truster), j (the trustee) wants that i will achieve its goal φ. Our definition of occurrent trust covers the situation in which a mother trusts a teacher to provide a good education to her son, even if the mother does not believe that the teacher's main motive to do that is to make the child's mother happy. According to the mother, the teacher's reason for providing a good education to children is just to ensure that they will be honest citizens at adult age. In this example, the mother's trust in the teacher is not based on the mother's belief about the teacher's adoption of her goal.

It is to be noted that other logicians have been interested in the issue of trust, and in particular in the relation between trust and security. Although there are several comprehensive logical models of security in which properties such as privacy, confidentiality, availability, integrity, authentication are modeled (e.g. [9]), there is still a pressing need for elaborating more general logical models of reasoning about trust. Indeed, logical models of trust have been focused almost exclusively on informational trust, i.e. trust in information sources [44, 38, 19, 18]. In these logics a certain agent is said to trust another agent if the former agent believes what the other agent says or that the information communicated to it by the other agent is reliable. Some authors have introduced trust in information sources as a primitive concept [44, 18] whereas other authors have reduced it to a particular kind of belief of the truster [38, 19]. An aspect of trust which is often neglected in existing logical models of trust is the motivational aspect. On the contrary, this aspect is central in our approach as well as in C&F's approach (see also [20] on the relationship between trust and motivational relevance). In the definitions of occurrent and dispositional trust proposed in Sections 4.1 and 4.2, i's trust in j necessarily involves a main and primary motivational component which is a goal of the truster. If i trusts j then, necessarily i trusts j because i has some goal (either occurrent or potential) and thinks that j has the right properties to ensure that such a goal will be achieved.

In this sense, our approach is different from the approach proposed by Jones [37] in which the motivational aspect is not considered to be necessary for defining trust, and trust is characterized only in terms of two beliefs of the truster: the truster's belief that a certain rule or regularity applies to the trustee (called "rule belief"), and the truster's belief that the rule or regularity is going to be followed by the trustee (called "conformity belief").[8] As stressed in Section 2, the goal component in the definition of trust is fundamental for C&F since it allows to distinguish trust from mere *thinking* and *foreseeing*.

[8]Note that this aspect of *regularity* emphasized by Jones is also explicit in our definition of dispositional trust in which the truster has to believe that *always*, if it will have the goal φ and the condition κ holds, the trustee will act in such a way that φ will be achieved. In this sense, the truster believes that there exists a regularity in the trustee's behavior.

5 From trust to reputation

In this section we parallel the preceding analysis of dispositional trust by an analysis of the reputation of a target agent j. First of all, we consider that reputation has *the same five ingredients* of dispositional trust: a target agent j that is the object of the reputation, an evaluating group of agents I, an action α of j, a goal φ of I with respect to which j's action is evaluated, and the relevant circumstances κ. Reputation therefore also takes the logical form of a 5-argument predicate

$$\texttt{Reput}(I,j,\alpha,\varphi,\kappa)$$

to be read "j has reputation in group I to do α with respect to φ in circumstances κ".

Is it always the case that the object of j's reputation is an action? Doesn't one just have the reputation of being a good physician or a good mechanics? We argue that even in these cases, reputation is about some set of actions in j's repertoire: j's reputation of being a good physician concerns a family of actions that j may perform, such as finding valid diagnoses, prescribing appropriate treatment, etc., and j's reputation of being a good mechanics is about j's actions of competently repairing tyres, brakes, motor, etc. Note that just as in the previous sections we are only considering intentional actions. We are aware that such a restriction to intentional actions is a limitation of our framework: as one of the reviewers pointed out, there are quite obvious some unintentional actions that might be the object of reputation; consider e.g. i's reputation of consenting when being drunk, or i's (bad) reputation of driving dangerously when being drunk.

The goal component is perhaps a bit more difficult to defend than in the case of trust: in a loose sense, reputation is about regular behavior of j that is known to group I, not involving any group goals, norms, standards or whatever. But we think that this is a matter of debate: it sounds odd to say that j has the reputation to take a coffee after lunch, because it does not matter for us whether j regularly drinks a cup of coffee after lunch or not, as long as this is not relevant for any of our goals. In contrast, j might be said to have the bad reputation to swallow whisky after dinner in case this action of j's is considered to violate some group standards (think e.g. of a group of muslims). In any case, such a kind of non-motivational reputation would be irrelevant in applications such as e-commerce. So we here consider a more restricted sense of reputation where α has to be relevant for the group goals.

Second, we shall argue that both trust and reputation can be defined from *the same components*, viz. j's capability, willingness, and opportunity. The main difference between trust and reputation is that the former is an individual belief of the truster, while the latter is a group belief of the evaluating agents. By "it is group belief that φ" we mean that the members of the group publicly accept that φ (or simply, that φ is public for the group). More precisely, while trust is an evaluation of a given target j by a certain agent i, reputation is a collective evaluation of a given target j by a group of agents I.

It remains to address the question what group beliefs are. We here adopt a notion of group belief which is close to Tuomela's notion of we-belief [62, 63]. First, we stress that *group belief* cannot be identified with the concept of common belief (or common

knowledge) that is familiar from theoretical computer science and philosophy [23, 43]. Indeed, while common belief in group I implies individual belief of every member of I, this should not be the case for the kind of evaluation of a target that is involved in reputation: else j's good reputation in group I would imply individual belief about the four properties in question, which is certainly too strong a link. Nevertheless, group belief should satisfy the following weaker property: the fact that the agents in I have a group belief that φ (or the agents in I publicly accept φ) implies that every agent in I believes that the agents in I have a group belief that φ. As this is supposed to be a logical principle and as our logic of belief has necessitation, it follows that it is also a common belief that φ is a group belief.

The following section is devoted to extend the logic of trust presented in Section 3 with operators of group belief that we have previously studied in [27] (see this work for more discussions on the concept of group belief and its logical properties; see also [49] for the related notion of acceptance). The extended framework will be applied to the formal characterization of the concept of reputation.

5.1 A logic for trust and reputation

In this section we extend the logic introduced in Section 3 with operators of the form $\texttt{Public}_I$, where I is a set of at least two agents. Operators of the form $\texttt{Public}_I$ allow to express that a certain fact φ is public for a group I (or that the group I believes that φ). We call $\mathcal{L}^+$ the extended logic.

Let

$$GR = \{A \subseteq AGT \ : \ card(A) \geq 2\}$$

be the set including all sets of at least two agents. The language of $\mathcal{L}^+$ is the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \vee \varphi \mid \texttt{G}\varphi \mid \texttt{After}_{i:\alpha}\varphi \mid \texttt{Does}_{i:\alpha}\varphi \mid \texttt{Bel}_i\varphi \mid \texttt{Choice}_i\varphi \mid \texttt{Public}_I\varphi$$

where p ranges over ATM, α ranges over ACT, i ranges over AGT and I ranges over GR. The formula $\texttt{Public}_I\varphi$ reads "it is public in group I that φ". Note that we might simplify the language and identify an individual belief $\texttt{Bel}_i\varphi$ with a singleton group's belief $\texttt{Public}_{\{i\}}\varphi$. However we refrain from this in order to avoid conceptual confusion.

The following abbreviation will be convenient:

$$\texttt{Poss}_I\varphi \stackrel{\text{def}}{=} \neg\texttt{Public}_I\neg\varphi$$

$\texttt{Poss}_I\varphi$ stands for "it is possible for group I that φ", or "group I does not exclude φ".

Frames of the logic $\mathcal{L}^+$ ($\mathcal{L}^+$-frames) are tuples $F^+ = \langle F, P\rangle$ where F is an $\mathcal{L}$-frame and P is defined as follows.

- $P : GR \longrightarrow W \times W$ maps every set I of at least two agents to a serial, transitive and Euclidean relation P_I between possible worlds in W.

The set $P_I(w)$ is the set of worlds w' which is compatible with I's group beliefs at w.

Models of the logic $\mathcal{L}^+$ ($\mathcal{L}^+$-models) are tuples $M^+ = \langle F^+, V\rangle$ defined as follows.

- F^+ is an $\mathcal{L}^+$-frame.
- $V : W \longrightarrow 2^{ATM}$ is a truth assignment which associates every world w with the set $V(w)$ of atomic formulas true in w.

The rule defining the truth condition of formula $\texttt{Public}_I\varphi$ is as usual:

- $M, w \models \texttt{Public}_I\varphi$ iff $M, w' \models \varphi$ for all $w' \in P_I(w)$

To the semantic constraints over $\mathcal{L}$-frames given in Section 3, we add the following. For every $w \in W$, and for every $I, J \in GR$ and $i \in AGT$ such that $i \in I$ and $J \subseteq I$:

S9 if $(w, w') \in P_J$ then $P_I(w) = P_I(w')$;

S10 if $(w, w') \in B_i$ then $P_I(w) = P_I(w')$;

S11 $P_I(w)$ is serial;

S12 if $(w, w') \in P_I$ then there exists $i \in I$ such that $(w', w') \in B_i$.

The axiomatization of $\mathcal{L}^+$ is given by all $\mathcal{L}$-principles given in Section 3 plus the following principles for the operators $\texttt{Public}_I$. For every $I, J \in GR$ and $i \in AGT$ such that $i \in I$ and $J \subseteq I$:

(**K**$_{\texttt{Public}}$)	$(\texttt{Public}_I\varphi \wedge \texttt{Public}_I(\varphi \rightarrow \psi)) \rightarrow \texttt{Public}_I\psi$
(**D**$_{\texttt{Public}}$)	$\neg(\texttt{Public}_I\varphi \wedge \texttt{Public}_I\neg\varphi)$
(**4**$_{\texttt{Public}_{I,J}}$)	$\texttt{Public}_I\varphi \rightarrow \texttt{Public}_J\texttt{Public}_I\varphi$
(**5**$_{\texttt{Public}_{I,J}}$)	$\neg\texttt{Public}_I\varphi \rightarrow \texttt{Public}_J\neg\texttt{Public}_I\varphi$
(**4**$_{\texttt{Public}_{I,i}}$)	$\texttt{Public}_I\varphi \rightarrow \texttt{Bel}_i\texttt{Public}_I\varphi$
(**5**$_{\texttt{Public}_{I,i}}$)	$\neg\texttt{Public}_I\varphi \rightarrow \texttt{Bel}_i\neg\texttt{Public}_I\varphi$
(**Unanim**)	$\texttt{Public}_I(\bigwedge_{i\in I} \texttt{Bel}_i\varphi \rightarrow \varphi)$
(**Nec**$_{\texttt{Public}}$)	From φ infer $\texttt{Public}_I\varphi$

Axioms **K**$_{\texttt{Public}}$ with rule of inference **Nec**$_{\texttt{Public}}$ are the principles of a minimal normal modal logic for every modal operator $\texttt{Public}_I$.

Axiom **D**$_{\texttt{Public}}$ expresses that public facts must be consistent, that is, φ and $\neg\varphi$ cannot be public in the same group I.

Axioms **4**$_{\texttt{Public}_{I,J}}$, **5**$_{\texttt{Public}_{I,J}}$, **4**$_{\texttt{Public}_{I,i}}$ and **5**$_{\texttt{Public}_{I,i}}$ express that a group of agents has always access to what is public (resp. not public) in its supergroups and that an agent has always access to what is public (resp. not public) in the group to which he belongs. According to axioms **4**$_{\texttt{Public}_{I,J}}$ and **4**$_{\texttt{Public}_{I,i}}$: if φ is public in I, then for all $J \subseteq I$ it is public in J that φ is public in I; if φ is public in I, then for all $i \in I$ agent i believes that φ is public in I. According to axioms **5**$_{\texttt{Public}_{I,J}}$ and **5**$_{\texttt{Public}_{I,i}}$: if φ is not public in I, then for all $J \subseteq I$ it is public in J that φ is not public in I; if φ is not public in I, then for all $i \in I$ agent i believes that φ is not public in I.

Axiom **Unanim** expresses a unanimity principle according to which it is public in every group I that if every member of I believes φ, then φ is the case. This axiom

describes the *bottom up* process leading from the beliefs of the agents in group I to I's group belief.

We call $\mathcal{L}^+$ the logic axiomatized by all $\mathcal{L}$-principles and the previous axioms and rules of inference for operators $\texttt{Public}_I$. We write $\vdash_{\mathcal{L}^+} \varphi$ if formula φ is a theorem of $\mathcal{L}^+$. We write $\models_{\mathcal{L}^+} \varphi$ if φ is *valid* in all $\mathcal{L}^+$-models. Finally, we say that φ is *satisfiable* if there exists a $\mathcal{L}^+$-model M and world w in M such that $M, w \models \varphi$.

We can prove that the logic $\mathcal{L}^+$ is *sound* and *complete* with respect to the class of $\mathcal{L}^+$-frames. Namely:

Theorem 5. *$\mathcal{L}^+$ is determined by the class of $\mathcal{L}^+$-frames.*

Proof. It is a routine task to check that the axioms of the logic $\mathcal{L}^+$ correspond one-to-one to their semantic counterparts on the frames. See proof of Theorem 1 for correspondences between the axioms of the logic $\mathcal{L}$ and the semantic constraints on $\mathcal{L}$ frames. Moreover, Axioms $\mathbf{4}_{\texttt{Public}_{I,J}}$ and $\mathbf{5}_{\texttt{Public}_{I,J}}$ together correspond to the constraint **S9**, whereas Axioms $\mathbf{4}_{\texttt{Public}_{I,i}}$ and $\mathbf{5}_{\texttt{Public}_{I,i}}$ together correspond to the constraint **S10**. Axiom $\mathbf{D}_{\texttt{Public}}$ corresponds to the constraint **S11**. Axiom **Unanim** corresponds to the constraint **S12**.

It is routine, too, to check that all axioms of the logic $\mathcal{L}^+$ are in the Sahlqvist class. This means that the axioms are all expressible as first-order conditions on frames and are complete with respect to the defined frames classes. □

5.2 A formal definition of reputation

We say that "j has reputation in group I to do α with respect to φ, in the circumstances κ" if and only if:

1. group I has the potential goal φ in the circumstances κ;
2. it is public for the group I that always, if every agent in I wants φ to be true and κ holds, then
   (a) j will be capable to do α;
   (b) j, by doing α, will ensure φ; and
   (c) j will intend to do α.

Before formalizing this notion of reputation, we introduce the following concept of "group I has the potential goal φ in the circumstances κ". We say that "group I has the potential goal φ in the circumstances κ" if and only if the agents in I do not publicly exclude that at some point in the future in which κ holds all of them will want φ. That is,

$$\texttt{PotGoal}_I^{\forall}(\varphi,\kappa) \stackrel{\text{def}}{=} \texttt{Poss}_I\texttt{F}^* \bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa)$$

One might object that this definition of potential group goal is quite strong since it requires that every agent in the group I will want φ to be true when κ holds. One might prefer a weaker notion. For instance, one might suppose "group I has the potential goal φ in the circumstances κ" if and only if the agents in I do not publicly exclude that

at some point in the future in which κ holds the majority of them will want φ. This alternative definition of potential group goal based on the concept of majority can be formally expressed as follows:

$$\texttt{PotGoal}_I^{Maj}(\varphi,\kappa) \stackrel{\text{def}}{=} \texttt{Poss}_I\texttt{F}^* \bigvee_{I' \subseteq I, |I'|>|I\setminus I'|}(\bigwedge_{i\in I'}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa)).$$

An even weaker notion of potential group goal requires that at least one agent in the group I will want φ to be true when κ holds. That is,

$$\texttt{PotGoal}_I^{\exists}(\varphi,\kappa) \stackrel{\text{def}}{=} \texttt{Poss}_I\texttt{F}^* \bigvee_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa).$$

In our view, there is no single definition of potential group goal. As shown in the domain of Social Choice Theory in which group preferences and goals are defined by means of an aggregation of individual preferences and goals, there is no single procedure for aggregating individual preferences and for building a corresponding definition of group preference. This is the reason why remain agnostic on this point. In the sequel we just consider the definition of potential group goal $\texttt{PotGoal}_I^{\forall}(\varphi,\kappa)$ based on universal quantification and we exploit it to define the concept of reputation. It is worth noting that alternative definitions of reputation based on different notions of potential group could be studied. We postpone this kind of analysis to future works.

We have now all necessary and sufficient ingredients to define the concept of reputation. As we have seen, the conditions $\texttt{Capable}_j(\alpha)$ and $\texttt{Intends}_j(\alpha)$ together are equivalent to $\texttt{Does}_{j:\alpha}\top$, and the definition of reputation can be simplified to:

$$\begin{aligned}\texttt{Reput}(I,j,\alpha,\varphi,\kappa) \stackrel{\text{def}}{=} &(\texttt{PotGoal}_I^{\forall}(\varphi,\kappa)\wedge \\ &\texttt{Public}_I\texttt{G}^*(\bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa) \rightarrow (\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi)))\end{aligned}$$

This means that, "agent j has the reputation in group I to ensure φ by doing α in the circumstances κ" if and only if, group I has the potential goal φ in the circumstances κ and, it is public in I that always, if everyone in I will want φ to be true and κ holds, then j will ensure φ by doing α.

EXAMPLE. Agent j is the employer of a certain company. Agent j has the reputation in the group I of employees to give incentives to them in order to increase their wages, under the condition that the company's income increases:

$$\texttt{Reput}(I,employer,giveIncentive_I,wageIncreased_I,incomeIncreased).$$

This means that the agents in I have the potential goal that their wages will increase, under the condition that the company's income increases:

$$\texttt{PotGoal}_I^{\forall}(wageIncreased_I,incomeIncreased).$$

Moreover, it is public for the group of employees that, whenever all employees will want their wages to increase under the condition that the company's income increases, the employer will give incentives so that wages will increase:

$$\begin{aligned}&\texttt{Public}_I\texttt{G}^*(\bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}wageIncreased_I \wedge incomeIncreased) \rightarrow \\ &(\texttt{Does}_{employer:giveIncentive_I}\top \wedge \texttt{After}_{employer:giveIncentive_I}wageIncreased_I)).\end{aligned}$$

The following theorems highlight some interesting properties of the notion of reputation.

Theorem 6. *Let $I, I' \in GR$, $\alpha \in ACT$ and $i, j \in AGT$ such that $I' \subseteq I$ and $i \in I$.*

Then:

(6a) $\vdash_{\mathcal{L}} (\texttt{Reput}(I,j,\alpha,\varphi,\kappa) \wedge \texttt{Public}_I \bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa)) \rightarrow \texttt{Public}_I\texttt{F}\varphi$

(6b) $\vdash_{\mathcal{L}} \texttt{Reput}(I,j,\alpha,\varphi,\kappa) \leftrightarrow \texttt{Public}_{I'}\texttt{Reput}(I,j,\alpha,\varphi,\kappa)$

(6c) $\vdash_{\mathcal{L}} \texttt{Reput}(I,j,\alpha,\varphi,\kappa) \leftrightarrow \texttt{Bel}_i\texttt{Reput}(I,j,\alpha,\varphi,\kappa)$

Proof. First, from the definition of $\texttt{G}^*$ we get
$\vdash_{\mathcal{L}^+} \texttt{Reput}(I,j,\alpha,\varphi,\kappa) \rightarrow \texttt{Public}_I(\bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi\wedge\kappa) \rightarrow (\texttt{Does}_{j:\alpha}\top\wedge\texttt{After}_{j:\alpha}\varphi))$.
With standard modal principles it then follows that
$\vdash_{\mathcal{L}^+} (\texttt{Reput}(I,j,\alpha,\varphi,\kappa) \wedge \texttt{Public}_I \bigwedge_{i\in I}(\texttt{Choice}_i\texttt{F}\varphi \wedge \kappa)) \rightarrow \texttt{Public}_I(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi)$.
Finally, $\vdash_{\mathcal{L}^+} \texttt{Public}_I(\texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi) \rightarrow \texttt{Public}_I\texttt{F}\varphi$ follows from $\vdash_{\mathcal{L}^+} \texttt{Does}_{j:\alpha}\top \wedge \texttt{After}_{j:\alpha}\varphi \rightarrow \texttt{F}\varphi$ as we have already seen in the proof of Theorem 2b.

Theorem 6b follows from the following theorems of $\mathcal{L}^+$: $\vdash_{\mathcal{L}^+} \texttt{Public}_I\varphi \leftrightarrow \texttt{Public}_{I'}\texttt{Public}_I\varphi$ and $\vdash_{\mathcal{L}^+} \neg\texttt{Public}_I\varphi \leftrightarrow \texttt{Public}_{I'}\neg\texttt{Public}_I\varphi$ for all $I' \subseteq I$ (these are provable by Axioms $\mathbf{4}_{\texttt{Public}_{I,J}}$ and $\mathbf{5}_{\texttt{Public}_{I,J}}$).

In order to prove Theorem 6c note first of all that due to Axioms $\mathbf{D}_{\texttt{Bel}}$, $\mathbf{4}_{\texttt{Public}_{I,i}}$ and $\mathbf{4}_{\texttt{Public}_{I,i}}$ we have for all $i \in I$:
$\vdash_{\mathcal{L}^+} \texttt{Public}_I\varphi \leftrightarrow \texttt{Bel}_i\texttt{Public}_I\varphi$ and $\vdash_{\mathcal{L}^+} \neg\texttt{Public}_I\varphi \leftrightarrow \texttt{Bel}_i\neg\texttt{Public}_I\varphi$.
Theorem 6c follows from these two theorems and the definition of $\texttt{Reput}(I,j,\alpha,\varphi,\kappa)$. □

According to Theorem 6a, if j has reputation in group I to do α with respect to φ in the circumstances κ, and it is public in I that everyone in I wants φ to be true and κ holds, then it is public in I that φ will be true at some point in the future. In this sense, the agents in I have a collective positive expectation to achieve their collective goal φ. Theorem 6b highlights that reputation is intrinsically a public notion, that is, for every subgroup I' of I, j has a certain reputation in group I if and only if it is public in I' that j has this reputation in I. Finally, according to Theorem 6c, if i is an agent in the group I then, j has a certain reputation in I if and only if i believes that j has this reputation in I. The left-to-right direction of this theorem will be better explained in Section 6 where Conte & Paolucci's concept of reputation will be discussed [16].

It is to be noted that in our logical model reputation does not necessarily imply trust, i.e., $\texttt{Reput}(I,j,\alpha,\varphi,\kappa) \rightarrow \texttt{DispTrust}(i,j,\alpha,\varphi,\kappa)$ is not valid. Indeed, what is public in a group I about a certain agent j does not necessarily correspond to what the agents in the group I believe about j. Moreover, the fact that every agent in I is disposed to trust agent j to do α with respect to φ in the circumstances κ does not necessarily imply that j has reputation in group I to do α with respect to φ in circumstances κ, i.e., $\bigwedge_{i\in I} \texttt{DispTrust}(i,j,\alpha,\varphi,\kappa) \rightarrow \texttt{Reput}(I,j,\alpha,\varphi,\kappa)$ is not valid.

6 Computational models of reputation

Several reputation and trust models have been proposed in the last years. Most of them propose or adopt a quite vague definition of trust and reputation and emphasize the

process of inferring reputation instead of the concept itself. In this section we present a comparison with some of these models, focusing on the (often implicit) properties of reputation. Among the proposals, we selected those that present a clear definition of reputation.

In order to provide a detailed comparison we use an analysis grid composed of eight criteria. They are the following.

Collective (Col). Reputation may be conceived either as an individual process or as a collective process. This criterion checks therefore whether this process is performed by each agent alone to construct its own local reputations or if it is the result of a collective process involving several agents to compute values and collect inputs (using, for instance, gossips).

Cognitive (Cog). Reputation can be represented in different ways, for instance, it can be represented by numerical values resulting from mathematical computation, by fuzzy sets or defined by means of mental states such as beliefs and goals. This criterion checks if a cognitive representation is used (e.g. using mental attitudes like beliefs and goals).

Evaluation group (Grp). We can consider reputation as a global estimation shared by the whole society of agents or as an estimation shared by a group of agents. In the latter case, different reputations are attached to an agent according to the group that judged it. Thus, the same agent can have a good reputation in a group and a bad reputation in another group. This criterion is satisfied if the reputation model allows the expression of group-dependent reputation (the term I in our definition).

Group goal (GG). Reputation may also depend on a given goal (as stressed throughout this article). The same agent can have different reputations with respect to different goals. If a reputation model considers this case, the criterion is satisfied.

Action ability (Abi). The agent's ability to perform an action has an impact on the way other agents will perceive its behaviors and then on its reputation. This criterion checks if the agents' ability is considered in the definition of reputation.

Action opportunity (Opp). The agent's opportunity to achieve a goal when performing an action has also an impact on the way other agents will perceive its behaviors and then on its reputation. This criterion checks if the agents' opportunities are considered in the definition of reputation.

Intention (Int). The definition of reputation may or not include the intention of the target agent. This intentional ingredient of reputation can be considered in two different ways: the agent's intention to perform an *action* α so that the goal of the group will be achieved or the agent's intention to achieve the *goal* of the group.

Context (Con). The three ingredients of trust/reputation (ability, opportunity and intention) can be context-dependent. For instance, an agent may have the opportunity to achieve a given goal by performing a given action only if some conditions are satisfied (the condition κ in our definition of reputation). This criterion checks if reputation considers such conditions.

Definition	Col	Cog	Grp	GG	Abi	Opp	Int	Con
ForTrust	yes	yes	yes	yes	yes	yes	action	yes
e-Bay	yes	no	no	no	no	no	no	no
Conte & Paolucci	yes	yes	yes	yes	no	no	goal	no
FIRE	yes	no	no	no	no	no	action	no
LIAR	yes	no	no	yes	no	no	no	yes
Regret	yes	no	yes	no	yes	no	no	no

Table 1: Properties of definitions of reputation

Table 1 contains a summary of the properties of other definitions of reputation proposed in the MAS domain using our definition and the above criteria as a base. In the table, our definition of reputation is identified by 'ForTrust'. In the sequel the other proposals are commented.

One of the most popular reputation model is probably that used in on-line systems like Amazon and e-Bay. The reputation of the participants is clearly the result of a collective process where users feedback the systems with evaluation of the transactions. The reputation is represented by a number that summarizes all the history of evaluations. This system however does not consider all the contextual reputation information. Moreover, it does not consider the reputation of a target with respect to a sub-group of participants. Finally, it does not consider group goal, the target's abilities, opportunities and intentions as dimensions of its reputation.

The definition of reputation presented by Conte & Paolucci in [16] shares fundamental properties with ours: reputation is a collective construction about the willingness of some agent towards some group goals. Conte & Paolucci distinguish image from reputation: the former is an evaluation of a given target shared by the agents in a group, the latter is a *voice* circulating in a group of agents about the target. Although the image of a target j in a group of agents I can be influenced, among other factors, by j's reputation in the group I, j's reputation in I does not necessarily coincide with its image. In Conte & Paolucci's view, it is not necessarily the case that, if j has a certain reputation in the group I relative to a certain property then every agent in I believes that j has this property. The reputation of j in I only implies that every agent in I believes that there is a voice about j which circulates in I. This aspect of reputation is captured by our formal definition based on the concept of group belief (see the left-to-right direction of Theorem 6c). Indeed, the fact that the agents in I have a group belief that φ (i.e. $\texttt{Public}_I\varphi$) does not necessarily imply i's belief that φ (i.e. $\texttt{Bel}_i\varphi$) when $i \in I$. On the contrary, the fact that the agents in I have a group belief that φ (i.e. $\texttt{Public}_I\varphi$) implies i's belief that the agents in I have a group belief that φ (i.e. $\texttt{Bel}_i\texttt{Public}_I\varphi$) when $i \in I$. In turn, in Conte & Paolucci's theory the following principle is supposed to be valid: if there is a voice about j which circulates in I then each agent in I believes that there is a voice about j which circulates in I. So, if we suppose that 'there is a voice about j which circulates in I' is equivalent to 'I have a group belief about j' then it logically follows from Conte & Paolucci's principle that: if there is a voice about j which circulates in I then each agent in I believes that I have a group belief about j.

Differently from our approach, in their notion of reputation Conte & Paolucci do not explicitly consider the target's ability and intention to perform a certain action α, and its opportunity to achieve a state of affairs φ by performing α (where φ is a goal of the group). Conte & Paolucci rather consider the target's intention to achieve the goal of the group.

For the FIRE model [36], trust is a "measurable level of the subjective probability with which an agent a assesses that another agent b will perform a particular action, both before a can monitor such action and in a context in which it affects its own action." If this value is not built from direct interactions with the target agent but from recommendations or opinions of others, it is called reputation. As in our proposal, reputation has a structure similar to trust, but with a collective dimension. This collective dimension, however, has not exactly the same meaning. In FIRE, the distinction between trust and reputation is given in terms of the kind of information source used; while own evaluations based on direct experiences lead to trust, opinions of others lead to reputation. Both, trust and reputation, are conceived from the point of view of the truster. There is no notion of evaluation of a target shared by a group of agents as proposed in our approach (i.e. what is public in a group about a given target). Thus, in FIRE, it is possible to state that an agent i assigns a reputation to agent j, but it is not possible to state that agent j has a certain reputation in the group of agents I.

The LIAR model uses reputation to control agents' behavior in a decentralized way [54]. As in FIRE, reputation is a subjective belief that is built from recommendations given by other agents. An agent gathers some recommendations from its neighbors about a given target when trying to achieve a given goal (that is usually a norm that must be respected in the society) in certain circumstances. As in our proposition, the use of the concept of group goal and context as ingredients of reputation are highlighted. However, we can draw the same conclusion than for FIRE: in LIAR it is not possible to state that an agent j has a certain reputation in a group of agents I.

Among direct experiences and opinions from other agents, in the Regret system [59], the reputation of an agent also considers its position in a group. More precisely, it considers the position of the target agent in a given sociogram. The reputation based on a sociogram is called 'neighborhood reputation'. While the two former dimensions of reputation are subjective to the agent (as it happens in FIRE), the latter dimension is similar to ours: the reputation of a target agent j is given in the context of a group I and is not reduced to the opinion of the members of I about j. If an agent collects all the opinions of the members of group I (witness reputation), it can not compute the neighborhood reputation from these opinions. The concepts of group goal, opportunity, intention and context are not considered in Regret. Concerning actions, Regret focuses on the evaluation of an agent's ability to perform a certain action.

7 Conclusion

We have presented in this article a modal logic of trust and reputation in which the concept of trust proposed by Castelfranchi & Falcone can be expressed, namely trust as the truster's belief about certain relevant properties of the trustee with respect to a given goal. We have also provided a refinement of C&F's concept by distinguishing

between two general categories of trust: occurrent trust and dispositional trust. In the second part of the article, we have dealt with the notion of reputation. In our approach, reputation is a collective evaluation of a given target j by a group of agents I (i.e. what is public in group I about j). It is structurally similar to dispositional trust but moves the basic concept of belief to a collective dimension of group belief.

Lorini and Demolombe [47] proposed a logic that is similar to ours, but having moreover a modal operator of obligation. Within that logic they have studied not only trust about actions, but also trust about dispositions: trust about motivational dispositions (alias intentions to be) and trust about normative dispositions. Within that framework they model several security properties and their relations with trust. In particular they define the concepts of trust about obedience and trust about honesty. One might parallel their approach by a straightforward extension of our logic by deontic operators.

Directions for future research are manifold. For instance, future works will be devoted to extend the logical model of trust and reputation presented in this article with a notion of *graded trust* based on a notion of *uncertain belief*. Indeed, in the present work we have only considered a notion of *binary trust* (i.e. either i trusts j or i does not trust j). Such a kind of extension will enable us to integrate the cognitive and qualitative analysis of trust presented in this article with a quantitative analysis and, to compare our approach with existing probabilistic approaches to trust (e.g. [40]). We have already started to study this topic in a recent work [46]. We also would like to give an account of bad reputation that is symmetric to that of good reputation of Section 5. At least at first glance, we can say that j has a bad reputation within the group I (w.r.t. j's action α and I's group goal φ) iff I has group goal φ and it is public in I that each time everyone in I wants φ to be true, j will prevent the agents in I from achieving their goal by doing α.

We also plan to implement and evaluate our proposal in an agent programming language where beliefs and goals are basic constructors (e.g. *Jason* [7] and 2APL [17]). The concepts of trust and reputation as presented in this article are more suitably integrated in such languages since they share the same background. An important issue to be investigated is the overall process of discovering the capabilities, opportunities, and intentions of the target agents based on their behavior. We plan to use the ART testbed scenario [25] to experiment and evaluate both the model and the implementation. Another topic of research is how to reify the operator $\texttt{Public}_I$ in such programming platforms. An initial proposal is to use the concept of shared artifacts as instruments to implement group beliefs. The infrastructure proposed in [58] will be used since it binds the above cited languages to artifacts. A preliminary work in that direction is presented in [35].

Acknowledgements

The work presented in this article is supported by the French Agence Nationale de la Recherche in the framework of the ForTrust[9] project (ANR-06-SETI-006).

We wish to thank the three reviewers of the NorMAS special issue: their careful and thorough reading and their detailed comments helped us to improve the paper and

[9] http://www.irit.fr/ForTrust/

correct some errors. We also would like to thank our colleagues of the ForTrust project Jonathan Ben-naim, Olivier Boissier, Cristiano Castelfranchi, Robert Demolombe, Rino Falcone, Dominique Longin and Laurent Perrussel for numerous discussions about the topic of the paper.

References

[1] K. Arrow. *The Limits of Organization*. Norton, New York, 1974.

[2] R. Audi. Intending. *Journal of Philosophy*, 70:387–402, 1973.

[3] A. Baier. Trust and antitrust. *Ethics*, 96:231–260, 1986.

[4] P. Battigalli and M. Dufwenberg. Dynamic psychological games. *Working paper, Mimeo, IGIER-Bocconi*, 2005.

[5] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[6] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, Cambridge, 2001.

[7] R. H. Bordini, J. F. Hübner, and M. Wooldridge. *Programming Multi-Agent Systems in AgentSpeak using Jason*. Wiley Series in Agent Technology. John Wiley & Sons, 2007.

[8] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.

[9] M. Burrows, M. Abadi, and R. M. Needham. A logic of authentication. *ACM Transactions on Computer Systems*, 8(1):18–36, 1990.

[10] C. Castelfranchi. Trust and reciprocity: misunderstandings. *International Review of Economics*, 55(1-2):45–63, 2008.

[11] 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.

[12] C. Castelfranchi and F. Paglieri. The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese*, 155:237–263, 2007.

[13] P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[14] J. Coleman. *Foundations of Social Theory*. Harvard University Press, Cambridge, 1990.

[15] R. Conte and C. Castelfranchi. *Cognitive and social action*. London University College of London Press, London, 1995.

[16] R. Conte and M. Paolucci. *Reputation in Artificial Societies. Social Beliefs for Social Order*. Kluwer, Boston, 2002.

[17] M. Dastani. 2APL: a practical agent programming language. *Autonomous Agent and Multi-Agent Systems*, 16:241–248, 2008.

[18] M. Dastani, A. Herzig, J. Hulstijn, and L. van der Torre. Inferring trust. In *Proceedings of the Fifth Workshop on Computational Logic in Multi-agent Systems (CLIMA V)*, volume 3487 of *LNCS*, pages 144–160. Springer-Verlag, 2004.

[19] R. Demolombe. To trust information sources: A proposal for a modal logic framework. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*. Kluwer, Dordrecht, 2001.

[20] M. Deutsch. Trust and suspicion. *The Journal of Conflict Resolution*, 2(4):265–279, 1958.

[21] M. Deutsch. *The resolution of conflict*. Yale University Press, New Haven, London, 1973.

[22] R. Fagin and J. Halpern. Reasoning about knowledge and probability. *Journal of the Association for Computing Machinery*, 41(2):340–367, 1994.

[23] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[24] 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.

[25] K. Fullam, T. B. Klos, G. Muller, J. Sabater-Mir, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. A specification of the Agent Reputation and Trust (ART) testbed: Experimentation and competition for trust in agent societies. In *Proceedings of the Fourth International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2005)*, pages 512–518. ACM Press, 2005.

[26] D. Gambetta. Can we trust trust? In D. Gambetta, editor, *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Basil Blackwell, Oxford, 2000.

[27] B. Gaudou, A. Herzig, and D. Longin. Grounding and the expresssion of belief. In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR'06)*, pages 221–229. AAAI Press, 2006.

[28] A. Goldman. *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs NJ, 1970.

[29] H. P. Grice. Intention and uncertainty. *Proceedings of the British Academy*, 57:263–279, 1971.

[30] J. Y. Halpern. *Reasoning about uncertainty*. MIT Press, Cambridge, 2003.

[31] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[32] J. Hintikka. *Knowledge and Belief*. Cornell University Press, New York, 1962.

[33] M. Hollis. *Trust whithin Reason*. Cambridge University Press, Cambridge, 1998.

[34] R. Holton. Deciding to trust, coming to believe. *Australian Journal of Philosophy*, 72(1):63–76, 1994.

[35] J. F. Hübner, O. Boissier, and L. Vercouter. Instrumenting multi-agent organisations with reputation artifacts. In *Proceedings of Coordination, Organizations, Institutions and Norms (COIN@AAAI)*, pages 17–24. AAAI Press, 2008.

[36] T. Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt. FIRE: An integrated trust and reputation model for open multi-agent systems. In *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, 2004.

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

[38] A. J. I. Jones and B. S. Firozabadi. On the characterization of a trusting agent: Aspects of a formal approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.

[39] C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In *Multi-Agent System Engineering: Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World (MAAMAW'99)*, pages 221–231. Springer, 1999.

[40] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[41] S. Kanger and H. Kanger. Rights and parliamentarism. *Theoria*, 6(2):85–115, 1966.

[42] R. M. Kramer. Trust and distrust in organizations: emerging perspectives, enduring questions. *Annual Review of Psychology*, 50:569–598, 1999.

[43] D. K. Lewis. *Convention: a philosophical study*. Harvard University Press, Cambridge, 1969.

[44] C. J. Liau. Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.

[45] L. Lindahl. Stig Kanger's theory of rights. In G. Holmström-Hintikka, S. Lindström, and R. Sliwinski, editors, *Collected Papers of Stig Kanger with Essays on his Life and Work*, volume 2, pages 151–171. Kluwer, Dordrecht, 2001.

[46] E. Lorini and R. Demolombe. From binary trust to graded trust in information sources: a logical perspective. In *Trust in Agent Societies 2008 (Selected papers)*, volume 5396 of *LNAI*. Springer-Verlag, 2008. to appear.

[47] E. Lorini and R. Demolombe. Trust and norms in the context of computer security. In *Proceedings of the Ninth International Conference on Deontic Logic in Computer Science (DEON'08)*, volume 5076 of *LNCS*, pages 50–64. Springer-Verlag, 2008.

[48] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77.

[49] E. Lorini, D. Longin, B. Gaudou, and A. Herzig. The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 2009.

[50] D. Makinson. On the formal representation of rights relations: remarks on the work of Stig Kanger and Lars Lindahl. *The Journal of Philosophical Logic*, 15:403–425, 1986.

[51] J. G. March and H. Simon. *Organizations*. John Wiley & Sons, 1958.

[52] S. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, University of Stirling, 1994.

[53] H. McCann. Settled objectives and rational constraints. *American Philosophical Quarterly*, 28:25–36, 1991.

[54] G. Muller and L. Vercouter. Decentralized monitoring of agent communications with a reputation model. In R. Falcone, S. Barber, J. Sabater-Mir, and M. Singh, editors, *Trusting Agents for Trusting Electronic Societies*, volume 3577 of *LNCS*, pages 144–161. Springer-Verlag, 2005.

[55] V. Pelligra. Under trusting eyes: the responsive nature of trust. In R. Sugden and B. Gui, editors, *Economics and Social Interaction: accounting for the interpersonal relations*. Cambridge University Press, Cambridge, 2005.

[56] S. D. Ramchurn, D. Huynh, and N. R. Jennings. Trust in multiagent systems. *The Knowledge Engineering Review*, 19(1):1 – 25, 2004.

[57] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pages 473–484. Morgan Kaufmann, 1991.

[58] A. Ricci, M. Piunti, L. D. Acay, R. H. Bordini, J. F. Hübner, and M. Dastani. Integrating heterogeneous agent programming platforms within artifact-based environments. In *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, pages 225–232. ACM Press, 2008.

[59] J. Sabater and C. Sierra. Reputation and social network analysis in multi-agent systems. In *First International Conference on Autonomous Agents and Multiagent systems (AAMAS-02)*, pages 475–482. ACM Press, 2002.

[60] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33 – 60, 2005.

[61] J. R. Searle. *The rediscovery of the mind*. MIT Press, Cambridge, 1992.

[62] R. Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, Stanford, 1995.

[63] R. Tuomela. *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge, 2002.

[64] J. Van Benthem. Correspondence theory. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 3, page 325408. Kluwer Academic Publishers, 2001. 2nd edition.

[65] G. H. Von Wright. *Norm and Action*. Routledge and Kegan, London, 1963.

[66] M. Witkowski, A. Artikis, and J. Pitt. Experiments in building experiental trust in a society of objective-trust based agents. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 111–132. Kluwer Academic Publishers, Dordrecht, 2001.