Prolegomena for a logic of trust and reputation*

Andreas Herzig[1], Emiliano Lorini[1], Jomi F. Hübner[2], Jonathan Ben-Naim[1], Olivier Boissier[2], Cristiano Castelfranchi[3], Robert Demolombe[1], Dominique Longin[1], Laurent Perrussel[1], and Laurent Vercouter[2]

[1] IRIT-LILaC, Toulouse, France
[2] ENS Mines Saint-Etienne, France
[3] ISTC-CNR, Rome, Italy

Abstract. Reputation and trust are useful instruments in multi-agent systems to evaluate agent behaviour. Most of the works on trust and reputation adopt a quantitative representation of these concepts. Trust and reputation are commonly simplified to a numerical representation loosing important properties of these concepts. The aim of this paper is therefore to provide a qualitative formal analysis of trust and reputation on the basis of cognitive primitives. The proposed formalization is strongly inspired by Castelfranchi and Falcone's model of social trust. The concepts of trust and reputation are built from the same bricks (goal, capability, power, and willingness) but in a different scope (individual belief vs. collective belief).

1 Introduction

The concepts of trust and reputation are 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 since a couple of years, and by now there exist manifold theoretical models and implemented systems.

One of the most prominent theoretical models is the cognitive model of trust by Castelfranchi and Falcone, henceforth abbreviated C&F [7, 8]. Their informal definition of trust is formulated as an individual belief about some properties of the trustee.

Our first aim is to give a more formal and more refined version of the C&F definition. Our strategy is top-down: we adopt C&F's reduction of trust in terms of the more primitive concepts of belief, goal, capability, power and willingness. We then link their notions of belief and goal to logics of belief and preference existing in the BDI literature. We finally step by step refine the last three concepts of capability, power and willingness by invoking other, yet more primitive concepts, namely action, preference and choice. To say that this definition provides a reduction means that we think that the concepts of belief, goal, capability, willingness and power are more primitive and better understood than that of trust. Based on this conceptualisation of trust, we then propose a definition of reputation that is structurally similar but moves the basic concepts of beliefs and goals to a collective dimension of group beliefs and goals. In fact,

* Work done within in the project "Social trust analysis and formalization" (ForTrust) that is supported by the French 'Agence Nationale de la Recherche' (ANR) within the SETIN call. Thanks are due to Sandrine Charbonnel for a careful reading of a previous version.

in our perspective trust and reputation have the same content: the properties of a given target. The only difference is that trust is an individual attitude (micro level), whereas reputation is a group attitude (macro level).

We should stress right from the start that there are several logics of belief, of goal, of capability, etc.: we thus do not provide a definitive logic of trust, but rather clarify the relevant concepts and offer a range of options of which we believe that they are good candidates for a formal basis of trust.

The contributions of the present paper are threefold. After recalling the C&F definition in Section 2 and distinguishing occurrent from dispositional trust in Section 3, we first give a formal logical analysis of occurrent trust (Section 4). Second, we give a formal logical analysis of dispositional trust (Section 5). Third, we provide a parallel definition of reputation in terms of a group belief about some properties of the target (Section 6). Finally, we evaluate to which extent the existing models of trust and reputation conform to these definitions (Section 7).

2 Informal definition of trust

Differently from other approaches [24, 38], in C&F's approach trust is not reduced to mere subjective probability which is updated in the light of direct interaction with the trustee and reputational information. Their model of social trust accounts for the truster's attribution process, that is, for the truster's ascription of internal properties to the trustee (capabilities, intention, dispositions, etc.) and for the truster's ascription of properties to the environment in which the trustee is going to act.

From this perspective, trust is not a simplistic notion. Two fundamental distinctions are introduced:

- 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 environmental 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 his 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 will be willing to act in a certain way).

According to C&F, trust has four ingredients: a truster i, a trustee j, an action α of j, and a goal φ of i. Throughout the paper, $i, j, \ldots$ denote agents (where usually i is the truster and j is the trustee), $\alpha, \beta, \ldots$ denote actions, and $\varphi, \psi, \ldots$ denote goals, and more generally logical formulas. C&F provide a definition of trust which is based on four primitive concepts: capability, intention, power, and goal. In their definition, "i trusts j to do α in order to achieve φ" if and only if:

1. i has the *goal* φ;
2. i believes that j is *capable* to do α;
3. i believes that j has the *power* to achieve φ by doing α; and
4. i believes that 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, (2) i believes that j is capable to send P, (3) that j's sending P will result in i possessing P, and (4) that j has the intention to send P.

C&F 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*.

The capability and the power concepts relate to the external attribution of i's trust, while the intention concept relates to the internal attribution.

Remark 1. In recent work Demolombe and Lorini [13] have generalized the motivational fourth component, by also considering norm-obedience. This allows them to distinguish goal-based trust from norm-based trust. In the sequel we shall stay with C&F's original definition, but will come back to this generalization in Section 6, because it allows to highlight the parallels between trust and reputation.

3 Distinguishing occurrent trust from dispositional trust

Let us have a closer look at the trustee's action α which is the object of trust. We distinguish two perspectives: in the first, the truster believes that the trustee is going to act *here and now*; in the second perspectives, the truster believes that the trustee is going to act *whenever some conditions are satisfied*.[4] 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.[5]

C&F define what may be called occurrent trust, in the sense that i trusts j *here and now* to perform α: when i trusts j's action (token) 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 his capability to perform α logically entail that he is indeed going to perform α (or at least that he is going to attempt to perform α, see e.g. [25]). We use the predicate $OccTrust$ to denote C&F's concept of occurrent trust. Its components are predicates of belief $Believes(i, \varphi)$, occurrent goal $OccGoal(i, \varphi)$, occurrent capability $OccCap(j, \alpha)$, occurrent power $OccPower(j, \alpha, \varphi)$, and occurrent intention $OccIntends(j, \alpha)$. We therefore have:

$$\begin{aligned} OccTrust(i, j, \alpha, \varphi) \stackrel{\texttt{def}}{=} \; & OccGoal(i, \varphi) \wedge \\ & Believes(i, OccCap(j, \alpha)) \wedge \\ & Believes(i, OccPower(j, \alpha, \varphi)) \wedge \\ & Believes(i, OccIntends(j, \alpha)) \end{aligned}$$

The predicates on the right hand side of the definition will be explained in Section 4.

[4] This relates to a standard distinction in philosophy of action: *action tokens* (alias concrete actions, action instances, 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 [18].

[5] This distinction between occurrent trust and dispositional trust relates to distinction between occurrent belief and dispositional belief employed by some philosophers (e.g. [31]).

Suppose that i currently does not have the goal to possess product P. According to the above 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 would send P under the appropriate circumstances (typically comprising j's belief that i has the goal to possess the product and has paid for it).

Generally speaking, when i trusts j about j's action type α then i believes that *henceforth*, j will perform α *under some conditions*. We use the predicate $DispTrust(i, j, \alpha, \varphi)$ to denote dispositional trust. Its components are predicates of belief $Believes(i, \varphi)$, potential goal $PotGoal(i, \varphi)$, conditional capability $CondCap(j, \alpha)$, conditional power $CondPower(j, \alpha, \varphi)$, and conditional intention $CondIntends(j, \alpha)$.

$$\begin{aligned} DispTrust(i, j, \alpha, \varphi) \stackrel{\texttt{def}}{=} \; & PotGoal(i, \varphi) \wedge \\ & Believes(i, CondCap(j, \alpha)) \wedge \\ & Believes(i, CondPower(j, \alpha, \varphi)) \wedge \\ & Believes(i, CondIntends(j, \alpha)) \end{aligned}$$

The three components $CondCap(j, \alpha)$, $CondPower(j, \alpha, \varphi)$ and $CondIntends(j, \alpha)$ in the definition of dispositional trust are conditioned versions of the corresponding components of occurrent trust. They describe the circumstances under which i expects that j has the capability to perform α, has the power to obtain φ via α, and is willing to perform α. The goal component $PotGoal(i, \varphi)$ is logically weaker than that of the definition of occurrent trust: $OccGoal(i, \varphi$ implies $PotGoal(i, \varphi)$, but not the other way round. Moreover, dispositional trust is more basic than occurrent trust: it is possible to infer the latter from the former, under some conditions. (We will show in Section 5 that this is indeed the case, and will provide an explanation of the predicates on the right hand side of the previous definition.) The converse does not hold.

The next two sections contain a detailed analysis of the components of C&F's definition, viz. of belief, goal, capability, intention and power. Let us stress that there is no such thing as 'the' logic of belief, 'the' logic of goal, 'the' logic of capability, etc., and that our aim is more modest, viz. to clarify the relevant concepts and their interrelation, and to offer a range of options. We do not systematically give exact mathematical definitions of these concepts, but rather provide pointers towards definitions in the literature that are more or less consensual, or at least prominent.

We are going to typographically distinguish the semi-formal predicates that we have used up to now (such as $OccGoal(i, \varphi)$ and $Believes(i, \psi)$) from the formal concepts that we will introduce in the rest of the paper. We shall use typewriter fonts for the latter, such as $\texttt{Bel}_i$ for the modal operator of belief.

4 Occurrent trust and its components

We define the components of occurrent trust: belief, occurrent goal, occurrent capability, occurrent power and occurrent intention. This will be done by means of modal operators of time, belief, preference, and action.

4.1 Temporal operators

All of our subsequent definitions require temporal operators $\texttt{Henceforth}$ and $\texttt{Eventually}$ because we have to be able to speak about the future. The formula $\texttt{Henceforth}\,\varphi$ reads "φ henceforth holds", and the formula $\texttt{Eventually}\,\varphi$ reads "φ eventually holds".

Logics of time are well-studied in modal logic and in theoretical computer science [36]. Their semantics is based on transition relations between possible states, i.e. (possibly infinite) automata.

4.2 Belief operators

The notion of belief is well-studied in the field of epistemic and doxastic logic since the early 1960s [22]. In its syntax, the fact that agent i believes that φ is written $\texttt{Bel}_i\,\varphi$, where $\texttt{Bel}_i$ is a so-called *modal operator of belief*, and φ is any formula. There is a large consensus in the literature that the logic of belief is the modal system KD45.

Modal operators of belief allows to conveniently express things. Consider for example the formula $\texttt{Bel}_i\,(\texttt{Bel}_j\,\varphi \vee \texttt{Bel}_j\,\neg\varphi)$: i believes that either j believes φ, or j believes the negation of φ. More concisely we may say that i believes that j knows[6] whether φ is true. To be able to express such things is crucial e.g. when we want to reason about epistemic trust, where the truster has beliefs about the trustee's competence.

We simply identify the belief predicate in the C&F definition as follows:

$$Believes(i,\varphi) \stackrel{\texttt{def}}{=} \texttt{Bel}_i\,\varphi$$

4.3 Goals as preferences about the future

Many different accounts of the concept of goal exist. Together with many approaches in the AI literature, we consider that goals are preferences about the future: i has goal φ means that among the futures possible for i, i prefers those where φ is eventually true. "I have the goal to possess product P" is identified with "I prefer futures where I possess P (over futures where I don't)".

Generally speaking preferences are partial preorders. We do not consider this here in order to keep things simple, and focus on binary preferences. Probably the most prominent account of binary preferences is Cohen and Levesque's [10]. It is our preferred one, too, one of the reasons being that it offers the advantage of neatly integrating belief and preference.[7] The fact that agent i prefers that φ is written $\texttt{Pref}_i\,\varphi$, where $\texttt{Pref}_i$ is a *modal operator of preference* and φ is any formula. The goal φ to be obtained might be a fact about the world, but also a belief or a goal, allowing us to express for example "i wants j to adopt goal φ" by $\texttt{Pref}_i\ \texttt{Eventually}\,\texttt{Pref}_j\,\varphi$, or "$i$ wants to know whether φ" by $\texttt{Pref}_i\ \texttt{Eventually}\,(\texttt{Bel}_i\,\varphi \vee \texttt{Bel}_i\,\neg\varphi)$.

[6] Here and elsewhere we somewhat sloppily use the term 'knows' instead of 'believes', because English does not provide a simple way to read the formula $\texttt{Bel}_j\,p \vee \texttt{Bel}_j\,\neg p$ in terms of belief.

[7] A second advantage is that it moreover supports the concept of intention, that is fundamental in the analysis of willingness, see Section 4.5.

To sum it up, we identify the C&F goal predicate as follows:

$$OccGoal(i, \varphi) \stackrel{\texttt{def}}{=} \texttt{Pref}_i\ \texttt{Eventually}\, \varphi$$

where $\texttt{Pref}_i$ is a modal operator of type KD45 satisfying introspection, as defined in [21].

Remark 2. As the formula $\texttt{Pref}_i\ \texttt{Eventually}\, \top$ is valid, our definition allows for tautologous goals. A standard way to avoid this is to add a negative condition, resulting in what has been called achievement goals [10]. That i has an achievement goal that φ is then identified with truth of $\texttt{Pref}_i\ \texttt{Eventually}\, \varphi \wedge \neg \texttt{Bel}_i\, \varphi$. We do not consider this here for simplicity (in particular we would have to change the definition of a potential goal in the definition of the $DispTrust$-predicate in Section 5).

4.4 Capability and power in dynamic logic

When we want to define what it means that an agent i is capable to do action α we have to look at logics of action. Basically, these logics model actions in terms of transition systems between states. There are two main traditions. First, logics of agency study the interplay between an agent and the outcomes he can bring about [4, 1, 28]. These logics thus abstract away from the particular action achieving the outcome. Second, propositional dynamic logic (PDL) studies the interplay between an action and its effects [20]. It has been shown for instance in [37] that dynamic logic is a suitable tool to characterize the concepts of capability and power. We focus on the latter, the main reason being that contrarily to logics of agency, there is a rich literature on its integration with logics of belief and goal (e.g. dynamic epistemic logics [2] or doxastic dynamic logic [32, 33]).

PDL distinguishes actions such as α from formulas such as φ and ψ, and its set of nonlogical symbols is made up of these two distinct categories. The formula $\texttt{After}_\alpha\, \varphi$ expresses that φ will be true after *every possible* execution of action α. Thus $\texttt{After}_\alpha\, \bot$ expresses that α is inexecutable.

Several extensions have been proposed in which an agent argument is added to the PDL operators. In such extensions, the formula $\texttt{After}_{i:\alpha}\, \varphi$ expresses that φ is true after every possible execution of action α by agent i. For every action α and agent i, $\texttt{After}_{i:\alpha}$ is a *modal operator of action*.

In this framework, the concept of (occurrent) capability is captured by the following abbreviation.

$$OccCap(j, \alpha) \stackrel{\texttt{def}}{=} \neg \texttt{After}_{j:\alpha}\, \bot$$

In this sense, we identify "j is capable to perform α" with "α can be executed by j".

The concept of (occurrent) power relates j's action α with i's goal φ: j's performance of α will make φ true 'here and now'.[8] Thus, when j has the power to achieve φ by doing α then, necessarily, if j does α then φ will obtain. We write this formally as:

$$OccPower(j, \alpha, \varphi) \stackrel{\texttt{def}}{=} \texttt{After}_{j:\alpha}\, \varphi$$

[8] In a more elaborate version it is sufficient that j's performance of α will raise the probability of φ. We do not consider this here in order to keep things simple.

Remark 3. We here do not consider the epistemic aspect of power. Indeed, as emphasized by [6, 3, 26], it is intrinsic to the concept of agent i's power to achieve a certain result φ by performing an action α that i is aware of his opportunity. For example, for a thief to have the power of opening a safe, the thief must know its combination. In this sense, defining j's power to achieve φ by doing α as the conjunction of $\texttt{After}_{j:\alpha}\,\varphi$ and $\texttt{Bel}_j\,\texttt{After}_{j:\alpha}\,\varphi$ would be more appropriate.

4.5 Intention-to-do as preferred action

In order to define intention, as in [13] we enrich the basic language of PDL with operators $\texttt{Does}_{i:\alpha}$ where a formula $\texttt{Does}_{i:\alpha}\,\varphi$ reads "agent i is going to do α and φ will be true afterwards". This allows to speak both about what an agent can do ($\neg\texttt{After}_{i:\alpha}\,\bot$) and about what an agent does ($\texttt{Does}_{i:\alpha}\,\top$). The relationships between operators of type $\texttt{After}_{i:\alpha}$ and operators of type $\texttt{Does}_{i:\alpha}$ is given by the principle

$$\texttt{Does}_{i:\alpha}\,\varphi \rightarrow \neg\texttt{After}_{i:\alpha}\,\neg\varphi$$

In particular $\texttt{Does}_{i:\alpha}\,\top \rightarrow \neg\texttt{After}_{i:\alpha}\,\bot$, expressing that if agent i is going to do α then i can do α.

Following Cohen and Levesque [10] we say that j has the occurrent intention to perform action α if and only if j prefers to perform action α 'here and now'. Such an intention is called present-directed (as opposed to future-directed intentions that take the form $\texttt{Pref}_j\ \texttt{Eventually}\,\texttt{Does}_{j:\alpha}\,\top$). Formally:

$$OccIntends(j,\alpha) \stackrel{\texttt{def}}{=} \texttt{Pref}_j\,\texttt{Does}_{j:\alpha}\,\top$$

where $\texttt{Pref}_j$ is Cohen&Levesque's preference operator as defined in Section 4.3, and $\texttt{Does}_{j:\alpha}$ is a dynamic logic operator as defined in Section 4.4.

4.6 Formal definition of occurrent trust

Summing up we get the following definition of occurrent trust:

$$\begin{aligned} OccTrust(i,j,\alpha,\varphi) \stackrel{\texttt{def}}{=}\ & \texttt{Pref}_i\ \texttt{Eventually}\,\varphi \wedge \\ & \texttt{Bel}_i\,\neg\texttt{After}_{j:\alpha}\,\bot \wedge \\ & \texttt{Bel}_i\,\texttt{After}_{j:\alpha}\,\varphi \wedge \\ & \texttt{Bel}_i\,\texttt{Pref}_j\,\texttt{Does}_{j:\alpha}\,\top \end{aligned}$$

5 Dispositional trust and its components

We now define the components of dispositional trust: belief, potential goal, conditional capability, conditional power and conditional intention. The latter three components are conditioned versions of the corresponding components of occurrent trust, that are prefixed by a temporal 'henceforth' operator.

5.1 Potential goals

A potential goal is a goal which the agent does not exclude to have one day as an occurrent goal:

$$PotGoal(i,\varphi) \stackrel{\texttt{def}}{=} \neg\texttt{Bel}_i\ \texttt{Henceforth}\,\neg OccGoal(i,\varphi)$$

5.2 Conditional capability

We define the predicate $CondCap(j, \alpha)$ as:

$$CondCap(j, \alpha) \stackrel{\texttt{def}}{=} \texttt{Henceforth}\,(\kappa_{OccCap(j,\alpha)} \rightarrow OccCap(j, \alpha))$$

The condition $\kappa_{OccCap(j,\alpha)}$ describes the circumstances under which the truster i expects that the trustee j has the (occurrent) capability to perform α.

The condition $\kappa_{OccCap(j,\alpha)}$ is what in the reasoning about actions field of AI is called the *executability precondition of* α: the condition under which it is factually ('physically') possible for j to perform α. In the case where α is the action of sending P, the executability precondition might be that j is not sick, that j knows some (possibly incorrect!) address of i, etc.

5.3 Conditional power

We define $CondPower(j, \alpha, \varphi)$ as:

$$CondPower(j, \alpha, \varphi) \stackrel{\texttt{def}}{=} \texttt{Henceforth}\,(\kappa_{OccPower(j,\alpha,\varphi)} \rightarrow OccPower(j, \alpha, \varphi))$$

The condition $\kappa_{OccPower(j,\alpha,\varphi)}$ describes the circumstances under which the truster i expects that the trustee j has the power to obtain φ via α.

The condition $\kappa_{OccPower(j,\alpha,\varphi)}$ is called the *effect precondition for* α *causing* φ in the reasoning about actions field: the condition under which performance of α results in φ being true. In the case of sending P such a condition might be that j's beliefs about i's address are correct, that the postal services are not on strike, etc.

5.4 Conditional intention

We define $CondIntends(j, \alpha)$ as:

$$CondIntends(j, \alpha) \stackrel{\texttt{def}}{=} \texttt{Henceforth}\,(\kappa_{OccIntends(j,\alpha)} \rightarrow OccIntends(j, \alpha))$$

The condition $\kappa_{OccIntends(j,\alpha)}$ describes the circumstances under which the truster i expects that the trustee j will intend to perform α. It is the most delicate one to specify. It describes under which conditions the trustee j is prepared to perform action α. This can also be called willingness.

There can be multiple sufficient reasons for the trustee j to be willing to perform action α, none of which is necessary. Agent j might be aware of i's goal φ or not,[9] j might expect rewards from i (or some authority) in case he performs α, or decide to perform α because he is obedient to some (formal or informal) norms, or by a pure altruistic attitude toward i. In the case of our running example, $\kappa_{OccIntends(j,\alpha)}$ might be that j believes that i has ordered and paid for the product P, that i has a proof for

[9] Suppose i believes j to be trustworthy in general concerning action α, i.e. $DispTrust(i, j, \alpha, \varphi)$, but actually knows that j believes that i has goal $\neg\varphi$: then i should not trust j to do α!

that, that non-delivery might be punished, etc. The truster i's beliefs about the trustee j's capability to perform α as well as j's power of achieving φ by that are typically also part of $\kappa_{OccIntends(j,\alpha)}$.[10]

In [13], some of these conditions are studied and formally defined, such as an agent's dispositional willingness to adopt the goals of other agents. In dispositional willigness, the condition $\kappa_{OccIntends(j,\alpha)}$ in the definition $CondIntends(j,\alpha)$ is expressed formally as follows.

$$\kappa_{OccIntends(j,\alpha)} \stackrel{\texttt{def}}{=} \texttt{Bel}_j\, \texttt{Pref}_i\, \texttt{Does}_{j:\alpha}\, \top$$

In this case, we say that j is willing to perform action α for i (or j has a dispositional willingness to perform action α for i) if and only if j is willing to perform action α under the condition that he believes that i wants him to do α.

A form of moral willingness or *obedience* is also studied. In this case, agent j is obedient to do α if and only if j is willing to perform action α under the condition that he believes to be obliged to perform action α. Formally:

$$\kappa_{OccIntends(j,\alpha)} \stackrel{\texttt{def}}{=} \texttt{Bel}_j\, \texttt{Oblig}\, \texttt{Does}_{j:\alpha}\, \top$$

where $\texttt{Oblig}$ is a modal operator of obligation, that may be considered to be the one of of Standard Deontic Logic [9].

5.5 Formal definition of dispositional trust

Summing things up we obtain the following definition of dispositional trust:

$$\begin{aligned} DispTrust(i,j,\alpha,\varphi) \stackrel{\texttt{def}}{=}\ & \neg\texttt{Bel}_i\ \texttt{Henceforth}\, \neg\texttt{Pref}_i\ \texttt{Eventually}\, \varphi \wedge \\ & \texttt{Bel}_i\ \texttt{Henceforth}\, (\kappa_{OccCap(j,\alpha)} \rightarrow \neg\texttt{After}_{j:\alpha}\, \bot) \wedge \\ & \texttt{Bel}_i\ \texttt{Henceforth}\, (\kappa_{OccPower(j,\alpha,\varphi)} \rightarrow \texttt{After}_{j:\alpha}\, \varphi) \wedge \\ & \texttt{Bel}_i\ \texttt{Henceforth}\, (\kappa_{OccIntends(j,\alpha)} \rightarrow \texttt{Pref}_j\, \texttt{Does}_{j:\alpha}\, \top) \end{aligned}$$

As we said, the conditions $\kappa_{OccCap(j,\alpha)}$ and $\kappa_{OccPower(j,\alpha,\varphi)}$ describe the executability and effect condition of α, while $\kappa_{OccIntends(j,\alpha)}$ describes the conditions under which j is willing to perform α for i. These conditions are nonlogical, and we cannot say anything more about them here.

Remark 4. According to the previous definition of dispositional trust, it is not guaranteed that it is possible for agent i that the three conditions $\kappa_{OccCap(j,\alpha)}$, $\kappa_{OccPower(j,\alpha,\varphi)}$ and $\kappa_{OccIntends(j,\alpha)}$ are jointly true when he will have the occurrent goal that φ. For the sake of simplicity in the presentation, we have not included this additional condition in the definition of dispositional trust, although in some cases it seems to be relevant. Nevertheless, it can be included by replacing the first element $PotGoal(i,\varphi)$ in the previous conjunction by: $\neg\texttt{Bel}_i\ \texttt{Henceforth}\, \neg(\texttt{Pref}_i\ \texttt{Eventually}\, \varphi \wedge \kappa_{OccCap(j,\alpha)} \wedge \kappa_{OccPower(j,\alpha,\varphi)} \wedge \kappa_{OccIntends(j,\alpha)})$.

As we have announced, occurrent trust can be inferred from dispositional trust under some conditions.

[10] Note that the trustee j's awareness of this is also relevant here: suppose that i knows that j is capable to perform α, while j wrongly believes that he is not.

Theorem 1. *The following formula is a valid for any multimodal logic where* $\texttt{Bel}_i$ *and* $\texttt{Henceforth}$ *are normal modal operators in Chellas's sense [9] and where the latter obeys the standard temporal logic principle* $\texttt{Henceforth}\,\varphi \rightarrow \varphi$*:*

$$\begin{array}{l}(\mathit{DispTrust}(i,j,\alpha,\varphi) \\ \quad \wedge \mathit{OccGoal}(i,\varphi) \\ \quad \wedge \mathit{Believes}(i,\kappa_{\mathit{OccCap}(j,\alpha)}) \\ \quad \wedge \mathit{Believes}(i,\kappa_{\mathit{OccPower}(j,\alpha,\varphi)}) \\ \quad \wedge \mathit{Believes}(i,\kappa_{\mathit{OccIntends}(j,\alpha)}) \;\;) \;\rightarrow\; \mathit{OccTrust}(i,j,\alpha,\varphi)\end{array}$$

6 Formal definition of reputation

In this section we parallel the preceding analysis of trust by an analysis of the reputation of a target agent j. First of all, we consider that reputation has *the same four ingredients*: an agent j that is the object of the reputation, an evaluating group on agents I, an action α of j, and a goal φ of I with respect to which j's action is evaluated. Reputation therefore also takes the logical form of a 4-argument predicate

$$\mathit{Rep}(I,j,\alpha,\varphi)$$

to be read "j has reputation in group I to do α in order to achieve φ".

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.

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. We think that this is a matter of debate. It sounds odd to say that j has the reputation to drink coffee after lunch: it does not matter for us whether j regularly drinks a cup of coffee after lunch or not at least as long as this is not relevant for any of our goals. In any case, such a kind of reputation would be of little interest in applications such as e-commerce. So we here consider a strict sense of reputation where α has to be relevant for the group goals.

Second, we argue that both trust and reputation can be defined from *the same concepts*, viz. j's capability, willingness, and power. 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.

It remains to address the question what group beliefs and group goals are.

First, we stress that *group belief* cannot be identified with the concept of common belief that is familiar from theoretical computer science and economy [15]. Indeed, while common belief of a group I implies individual belief by every member of I, this should not be the case for the kind of group belief 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. We instead adopt Tuomela's definition of we-belief [34, 35], that was cast in a formal logic in [16, 17]. Group belief (noted $\texttt{GroupBelief}_I$) basically has the same properties as individual and common belief,

except that the infinitary construction of common belief[11] is abandoned. Group belief satisfies the following positive introspection property: the fact that the agents in I have a group belief that φ (i.e. $\texttt{GroupBelief}_I\,\varphi$) implies i's belief that the agents in I have a group belief that φ (i.e. $\texttt{Bel}_i\,\texttt{GroupBelief}_I\,\varphi$) when $i \in I$.

Let us turn now to *group goals*. We would like to understand them in a broad sense, including norms, standards, values and any other property that is in some sense ideal for the group. Therefore group goals are weaker than joint goals and joint intentions [14, 19]. Thus, the expression "the group of agents I has the (group) goal that φ" means for us "according to the group of agents I it *should* to be the case that φ". There is a lot of work about the relation between individual goals and group goals. For instance, we might capture a notion of group goal by means of an aggregation of the individual preferences, as done in classical social choice theory. It appears that we do not need such sophistications at the present stage. We therefore remain agnostic on this point, and do not detail any particular relationship between group preference $\texttt{GroupPref}_{\{i,j\}}\,\varphi$ on the one hand, and the individual preferences $\texttt{Pref}_i\,\varphi$ and $\texttt{Pref}_j\,\varphi$ on the other.

Summing things up, we define the predicate $Rep(I, j, \alpha, \varphi)$ as the conjunction of a goal of I and three beliefs of I about j:

$$\begin{aligned} Rep(I, j, \alpha, \varphi) \stackrel{\texttt{def}}{=}\ & PotGoal(I, \varphi) \wedge \\ & \texttt{GroupBelief}_I\, CondCap(j, \alpha) \wedge \\ & \texttt{GroupBelief}_I\, CondPower(j, \alpha, \varphi) \wedge \\ & \texttt{GroupBelief}_I\, CondIntends(j, \alpha) \end{aligned}$$

where $PotGoal(I, \varphi)$ is defined similarly to $PotGoal(i, \varphi)$ by:

$$PotGoal(I, \varphi) \stackrel{\texttt{def}}{=} \neg\texttt{GroupBelief}_I\ \texttt{Henceforth}\ \neg\texttt{GroupPref}_I\ \texttt{Eventually}\,\varphi$$

and the predicates $CondCap(j, \alpha)$, $CondPower(j, \alpha, \varphi)$ and $CondIntends(j, \alpha)$ are defined as in Section 5.

7 Comparison of reputation models

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. 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. The Table 1 contains a summary of the properties of other definitions of reputation using our definition as the base. A brief explanation of each criterion follows.

Collective (Col): whether the reputation is the result of a collective/global process.
Cognitive (Cog): if mental states like beliefs are used to define the reputation.
Evaluation group (Grp): if a target agent can have different reputations for different groups (the term I in our definition).

[11] Either by means of an induction axiom, or by means of the infinite conjunction $\texttt{CommonBelief}_{i,j}\,\varphi \stackrel{\texttt{def}}{=} \texttt{Bel}_i\,\varphi \wedge \texttt{Bel}_j\,\varphi \wedge \texttt{Bel}_i\,\texttt{Bel}_j\,\varphi \wedge \texttt{Bel}_j\,\texttt{Bel}_i\,\varphi \wedge \ldots$

Definition	Col	Cog	Grp	GG	Cap	Pow	Int	Con
ForTrust	yes	yes	yes	yes	yes	yes	action	yes
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
e-Bay	yes	no	no	no	no	no	no	no

Table 1. Properties of definitions of reputation

Group goal (GG): if the reputation can be assigned to a goal and thus the same agent can have different reputations for different goals (the term φ in our definition).

Action capability (Cap): whether the definition considers the capability of the agent for an action (the $CondCap$ predicate).

Action power (Pow): whether the definition considers the power of the agent to achieve the group goal when performing the action (the $CondPower$ predicate).

Intention (Int): whether the definition considers the agent's intentions to perform the action or goal for the group (the $CondIntends$ predicate).

Condition (Con): whether the reputation considers the dispositional conditions for the actions (the κ function).

The definition of reputation presented in Conte & Paolucci in [11] shares important 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. Indeed, the fact that the agents in I have a group belief that φ (i.e. $\texttt{GroupBelief}_I\,\varphi$) 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. $\texttt{GroupBelief}_I\,\varphi$) implies i's belief that the agents in I have a group belief that φ (i.e. $\texttt{Bel}_i\,\texttt{GroupBelief}_I\,\varphi$) when $i \in I$.

Besides the formalisation in relation to trust, our contribution introduces the actions of the target agent into the definition of reputation. This is more precisely defined in terms of the capability, power, and intention for some action α. Another difference is the intentional part of reputation which is cited in their definition as being related to a goal whereas in our definition the intention is assigned to an action.

For the FIRE model [23], 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 based on direct interactions with the target agent, but based on a

collection of experiences of other agents that interacted with the target agent, it is called reputation. In the FIRE model reputation is not analyzed at a collective level. Indeed, there is no notion of evaluation *shared* by the whole group. Differently from FIRE, in our approach an agent can distinguish in its belief base its own evaluation of the target (in our terms, its trust) from the evaluation of the target shared by the group (in our terms, its reputation).

The LIAR model uses reputation to control agents' behaviour in a decentralised way [27]. As in FIRE, reputation is an evaluation about a certain target agent which is built on the basis of the reports of other agents. An agent gathers some recommendations from its neighbours about a given target when trying to achieve a given goal (that is usually a norm that must be respected in the society) in a given context. Regarding our definition of reputation, this context corresponds to the dispositional condition of the target agent.

Among direct experiences and opinions from others, in the Regret system [30], the reputation of an agent also considers its position in a group, a position given by a sociogram. The reputation based on a sociogram is called 'neighbourhood reputation'. While the two former dimensions of reputation are subjective to the agent (as for 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 members of I (witness reputation), it cannot compute the neighbourhood reputation from these opinions. As the group goal is not considered, j's power needs not be considered either. Conditions and intentions are also not taken into account. Regarding the actions, the Regret model focuses on the external attribution of trust, i.e. the evaluation of an agent's capability to perform some action.

8 Conclusion

We have provided a logical analysis of the concepts of trust and reputation. Our first contribution is the formalization of Castelfranchi & Falcone's concept of occurrent trust within a combination of existing logics for MAS. Our second contribution is the identification of the concept of dispositional trust, that we have contrasted with C&F's occurrent trust. Our third contribution is a formal definition of reputation that is structurally similar to the definition of dispositional trust but moves from individual beliefs and goals to group beliefs and goals. According to our definitions, trust and reputation have the same content: some properties of a given target (capability, power, willingness). The only difference is that trust is an individual attitude (micro level), whereas reputation is a group attitude (macro level). Trust is an individual belief about some properties of a given target which are relevant for a goal of the truster, whereas reputation is a group belief about the same properties of the target which are relevant for a group goal. We have concluded by comparing our approach with existing models of trust and reputation.

Our future work will be devoted to present a semantics and an axiomatisation of the operators of belief, goal, action, group belief, group goal presented in this paper and of their interactions. We will also study the properties of the formal definitions of trust and reputation and show how reputation can be exploited by an agent in order to build

its trust in a given target. Special attention will be given to the issue of group goals in order to clarify their relationships with individual goals.

We also plan to implement our proposal in an agent programming language where beliefs and goals are basic constructors (e.g. *Jason* [5] or 2APL [12]). To manage group beliefs, we will evaluate the use of shared artefacts that are available in the agents' environment. These artefacts will then build and store the group beliefs required for reputation. The infrastructure proposed in [29] will be used since it binds the above cited languages to artefacts.

References

1. R. Alur, T. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49:672–713, 2002.
2. A. Baltag and L. S. Moss. Logics for epistemic programs. *Synthese*, 139 (2):165–224, 2004.
3. B. Barnes. *The Nature of Power*. Polity Press, Cambridge, 1988.
4. N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.
5. R. H. Bordini, J. F. Hübner, and M. Wooldrige. *Programming Multi-Agent Systems in AgentSpeak using Jason*. Wiley Series in Agent Technology. John Wiley & Sons, 2007.
6. C. Castelfranchi. The micro-macro constitution of power. *Protosociology*, 18-19, 2003.
7. C. Castelfranchi and R. Falcone. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proc. of ICMAS'98*, pages 72–79, 1998.
8. C. Castelfranchi and R. Falcone. Social trust: A cognitive approach. In C. Castelfranchi and Y. H. Tan, editors, *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer Academic Publishers, Dordrecht, 2001.
9. B. F. Chellas. *Modal logic: an introduction*. Cambridge University Press, Cambridge, 1980.
10. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
11. R. Conte and M. Paolucci. *Reputation in Artificial Societies. Social Beliefs for Social Order*. Kluwer, Boston, 2002.
12. M. Dastani, D. Hobo, and J.-J. C. Meyer. Practical extensions in agent programming languages. In E. H. Durfee, M. Yokoo, M. N. Huhns, and O. Shehory, editors, *6th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007), Honolulu, Hawaii, USA, May 14-18, 2007*, page 138. ACM Press, 2007.
13. R. Demolombe and E. Lorini. Trust and norms in the context of computer security: a logical formalization. In R. van der Meyden and L. van der Torre, editors, *Proceedings of the Ninth International Conference on Deontic Logic in Computer Science (DEON'08)*, LNAI. Springer-Verlag, forthcoming.
14. B. Dunin-Keplicz and R. Verbrugge. Collective intentions. *Fundamenta Informaticae*, 51(3):271–295, 2002.
15. R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.
16. B. Gaudou, A. Herzig, and D. Longin. Grounding and the expresssion of belief. In P. Doherty, J. Mylopoulos, and C. Welty, editors, *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR'06)*, pages 221–229, Lake District, GB, May 2006. AAAI Press.
17. B. Gaudou, D. Longin, E. Lorini, and L. Tummolini. Anchoring institutions in agents' attitudes: towards a logical framework for autonomous MAS. In *Proceedings of 7th International Joint Conference on Autonomous Agents in Multi-Agent Systems (AAMAS'08)*, to appear.

18. A. Goldman. *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs NJ, 1970.
19. B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.
20. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.
21. A. Herzig and D. Longin. C&L intention revisited. In *Proceedings KR04*, 2-5 june 2004.
22. J. Hintikka. *Knowledge and Belief*. Cornell University Press, New York, 1962.
23. T. D. Huynh, N. R. Jennings, and N. 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.
24. C. M. Jonker and J. Treur. Formal analysis of models for the dynamics of trust based on experiences. In F. J. Garijo and M. Boman, editors, *Multi-Agent System Engineering: Proceedings of the 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer Verlag, Berlin, 1999.
25. E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, forthcoming.
26. E. Lorini, A. Herzig, J. Broersen, and N. Troquard. Grounding power on actions and mental attitudes. In *Proceedings of FAMAS 2007*, pages 203–220, 2007.
27. G. Muller and L. Vercouter. *Trusting Agents for Trusting Electronic Societies*, volume 3577/2005 of *Lecture Notes in Computer Sciences*, chapter Decentralized Monitoring of Agent Communications with a Reputation Model, pages pp. 144–161. Springer, 2005.
28. M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
29. 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 *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008*. ACM Press, 2008.
30. 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, 2002.
31. J. R. Searle. *The rediscovery of the mind*. MIT Press, Cambridge, 1992.
32. K. Segerberg. Getting started: Beginnings in the logic of action. *Studia Logica*, 51(3-4):347–378, 1992.
33. K. Segerberg. Belief revision from the point of view of doxastic logic. *Logic Journal of IGPL*, 3(4):535–553, 1995.
34. R. Tuomela. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press, Stanford, 1995.
35. R. Tuomela. *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge, 2002.
36. J. van Benthem. *The logic of time*. Reidel Publishers, 1991. (first edition in 1983).
37. B. Van Linder, van der Hoek, and J.-J. C. W., Meyer. Formalising abilities and opportunities. *Fundamenta Informaticae*, 34:53–101, 1998.
38. 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.