Computer-Mediated Trust in Self-interested Expert Recommendations

Jonathan Ben-Naim, Jean-François Bonnefon, Andreas Herzig, Sylvie Leblois, Emiliano Lorini

Computer-Mediated Trust in Self-Interested Expert Recommendations

Jonathan Ben-Naim
CNRS and Université de Toulouse

Jean-François Bonnefon[1]
CNRS and Université de Toulouse

Andreas Herzig
CNRS and Université de Toulouse

Sylvie Leblois
Université de Toulouse

Emiliano Lorini
Université de Toulouse

May 27, 2009

[1]To whom correspondence should be addressed: Jean-François Bonnefon, CLLE-LTC, Maison de la Recherche, 5 al. Antonio Machado, 31058 Toulouse Cedex 9 France (email: bonnefon@univ-tlse2.fr).

Abstract

Important decisions are often based on a distributed process of information processing, from a knowledge base that is itself distributed among agents. The simplest such situation is that where a decision-maker seeks the recommendations of experts. Because experts may have vested interests in the consequences of their recommendations, decision-makers usually seek the advice of experts they trust. Trust, however, is a commodity that is usually built through repeated face time and social interaction, and thus cannot easily be built in a global world where we have immediate internet access to a vast pool of experts. In this article, we integrate findings from experimental psychology and formal tools from Artificial Intelligence to offer a preliminary roadmap for solving the problem of trust in this computer-mediated environment. We conclude the article by considering a diverse array of extended applications of such a solution.

1 Introduction

Important decisions are rarely made in isolation. Even when a single agent has the final say about what is to be done, the knowledge and information processing relevant to a complex decision are often distributed among several agents. Typically, one agent (the decision-maker) relies on one or several other agents (the experts) to provide recommendations based on their knowledge and know-how about the problem at hand.

The problem with experts, though, is that they may well have vested interests in the consequences of their recommendations. Think about investing. All of us who are not investment savy might want to get some expert recommendations about what to do with our savings. Sometimes, our banker is willing to provide such recommendations. We are likely to take these recommendations with a grain of salt, though, because we are aware that the banker may have vested interests in pushing some specific financial products. We are facing a dilemma between our need for expert recommendation and the potentially self-interested character of the recommendations we can get from experts, who have vested interests in the decision we are going to make from their recommendation.

Our need for expertise thus makes us the potential targets of deception from self-interested experts. The traditional solution to this dilemma is to seek the recommendations of these experts and only these experts whom we endow with our *trust.* Trust is a multidimensional concept that has been informally defined as, e.g., "the expectation that the person is both competent and reliable, and will keep your best interest in mind" [2], or, quite similarly, as a combined judgment of the integrity, ability, and benevolence of an individual [23].

Trust is a commodity that is often built through repeated social interactions [9, 17]. Not only do people trust other people as a function of their interpersonal history, but even the most subtle aspects of face-to-face interaction can contribute to judgments of trustworthiness. For example, people are more ready to trust interaction partners who mimic their behavioral mannerisms [22]. Whether or not this is a sensible way to endow someone with trust is, of course, a debatable question. The point is, though, that behavioral mimicry requires face-to-face interaction, and that, generally speaking, feelings of trust commonly require a history of social interaction with the person whom is to be trusted.

This solution to the problem of trust is adapted to a small world where experts on a given topic are few and personally known to the decision maker. However, in our global village, an inexhaustible pool of experts on just any given topic is always just one mouse clik away from us. Whatever our concern

is, the internet gives us a fast and convenient access to a vast number of experts. That would be good news, if only we knew which experts we could trust. The traditional solution to the problem of trust (repeated face time and social interaction) is no longer available in our global cognitive world.

In this article, we consider the problem of seeking expert advice through a web-based platform, where users are declared experts in various domains. We offer a list of suggestions for solving the problem of trust in this environment. The power of our approach resides in its multidisciplinarity, as we combine the cognitive insignts of psychology to the formal methods of artificial intelligence to reach an integrated perspective on our problem. In the solution that we envision, regular users alternatively play the role of advisor or advisee in their interactions, depending on whom is in possession of the expert knowledge required by the situation. After each interaction, advisees have the possibility to appraise their advisor on the various dimensions that form the multifaceted concept of trust. The platform keeps a memory of these appraisals from which it can extract an aggregated, global index of the trustworthiness of any user, or decompose this global index into sub-indices corresponding to the various components of trust. Any new or regular user can thus attain a computer-mediated judgment of the extent to which any expert on the platform is to be trusted, or seek an expert whose detailed characteristics are optimally balanced to serve their needs.

In the rest of this article, we give a more detailed characterization of our problem, and we address in turn the various ingredients we need to sketch a solution. Section 2 defines the problem of attaining a computer-mediated, complex judgment of trust within a multi-user, web-based platform of potentially self-interested experts. Section 3 reviews experimental findings and psychological insights into the components of trust, their socio-cognitive antecedents, and their behavioral consequences. Section 4 builds on these materials and on the current state of the art in artificial intelligence to sketch a formal solution to our problem, which integrates the psychological constraints previously identified. Finally, Section 5 considers various extended applications of our suggestion for computer-mediated trust.

2 Problem Specification

Let us imagine that you came in possession of a banjo, which you would like to sell, but whose monetary value you have no idea of. One option would be to go to the closest (and probably the only) banjo store you know of, and to ask the owner to appraise your banjo. The problem, though, is that the owner is not only the person whom you can ask about the value of your

banjo, but also the person you are likely to sell the banjo to. Not knowing whether you can trust the owner not to take advantage of the situation, you turn to a web-based platform for musical instruments amateurs, where you are likely to find plenty of users who can appraise a banjo, and plenty of potential banjo buyers. Your trust problem, though, is just demultiplicated, because these are likely to be broadly the same persons. The fact that you now have an abundance of experts you might solicit is no improvement over your previous situation, because you do not have the time, the resources, or the motivation to engage in repeated social interactions with all these people in order to find out who you can trust.

We believe that the platform should offer a solution to achieve the same results as repeated social interaction. It should provide you with the basic parameters that form the building blocks of trust, as well as some index of the extent to which you can trust your potential advisors. We believe this service can be achieved by formalizing the notion of trust, and taking advantage of the history of the advisor-advisee interactions on the platform.

Not every user of the platform is an expert of everything. To continue our musical instruments exemple, some users may declare expertise in appraising banjos, whilst others may declare expertise in appraising cellos. Thus, depending on the situation, a given user may be in a position to give expert advice, or to receive it. Now consider that everytime a user x receives expert advice from another user y, x is given the opportunity to appraise this advice on all the dimensions that the complex notion of trust is known to encompass. The platform records this interaction as a tuple $R_{xy} < r_1, \ldots, r_n >$, where $r_1, \ldots, r_n$ are the appraisals given by x about the recommendation of y on the various dimensions of trust. Soon enough, the platform should be in a position to answer a request about the trustworthiness of agent y, by aggregating the information contained in the tuples expressed about y.

A number of problems must be solved to achieve such a result. First, we need to decide on the exact nature of the appraisals $r_1, \ldots, r_n$. Then, we need to decide on the way these ratings should be aggregated, both at the individual level and at the collective level. Finally, we need a formal characterization of all the components of trust and of the properties one can used to reason about them, in order to generalize our solution to environments where artificial agents interact with human agents, or among themselves. Solving these problems requires a multidisciplinary approach, drawing on experimental psychology as well as artificial intelligence methods. We now consider in turn the insights given by these two disciplines.

3 Psychological Treatment

Various definitions of trust can be found in the psychological literature. Some authors define trust mostly in terms of its behavioral consequences, e.g., 'Trust is the extent to which a person is confident in, and willing to act on the basis of, the words, actions, and decisions of another' [24], or trust is 'the willingness to accept vulnerability based upon positive expectations about another's behavior' [27]. Early structural perspective on trust distinguished between trust based on cognition and trust based on affect [16, 18, 26]. 'Cognitive' trust is based on explicit knowledge and 'good reasons' to think that a person is reliable or dependable. 'Affective' trust is based on an emotional bond between individuals. Clearly, just as behavioral mimicry, emotional bonds are not within the scope of our application. We should thus focus on that sort of trust which is based on explicit knowledge and deliberative thought.

Idealilly suited for our purpose is the suggestion that trustworthiness is a three-dimensional attribute composed of competence, benevolence, and integrity [2, 23]. Competence reflects the ability of a person with respect to the task at hand. Benevolence reflects a positive attitude towards the truster, and a genuine concern for the truster's interests. Integrity reflects the adherence of the trustee to an appropriate set of ethical principles. Let us now consider in turn these three components of trust, and their potential importance in situation where advice is given.

3.1 Competence

Many studies have investigated the influence of an advisor's perceived competence on the uptake of her recommendations. Perhaps unsurprisingly, these studies concur that the recommendation of an advisor is more influential when her perceived expertise is greater. Interestingly, people seem ready to accept claims of expertise at face value, even in experimental situations where the quality of the offered 'expert' advice is actually weak [12]. While it is clear why people seek the advice of individuals they believe to be more competent than they are, we note that people are sometimes ready to seek the advice of individuals they believe to be *less* competent than they are; in particular, when the stakes of the decision are serious enough that they want to share the responsibility for the decision, whatever the relative expertise of their advisor [12].

People appear to use a variety of cues to appraise the expertise or competence of an advisor. For example, advisors who express high confidence in their recommendation are perceived as more competent, and their recom-

mendation is given more weight by the decision maker [31, 32]. Likewise, advisors who give very precise recommendations (as opposed to vague estimates) are perceived as more competent, and, again, their recommendation is given more weight by the decision maker [34]. All other things being equal, these strategies do appear to increase the quality of the decision making, for there seems to be an ecologically valid correlation between expertise, confidence, and precision [31, 32, 35]. Then again, these studies did not control for the possibility that the advisor has vested interests in the decision of the advisee; and a self-interested advisor may well express a very precise recommendation with great confidence, only to better serve her own interests.

Finally, a reputation for expertise is hard to build, but rapidly destroyed [30, 34]. Many useful recommendations are required before one is trusted as a competent advisor, but only a few average or bad recommendations are enough to lose that reputation. This phenomenon can be related to the more general *negativity bias* in impression formation [15, 29]. The negativity bias refers to the greater weight we attribute to negative behaviors when inferring personality traits: For example, fewer negative behaviors are needed to infer a negative trait, compared with the number of positive behaviors we need to infer a positive trait. The negativity bias, and its specific consequences for the dynamics of trust, can be conceived as a safeguard for a species that exhibits a strong tendency to spontaneous cooperation, ensuring that untrustworthy partners are quickly detected and unprofitable cooperation promptly forsaken.

3.2 Benevolence

Whenever a conflict of interest is possible, and even when it is not, people are concerned about the degree to which their advisors really care about their interests. A benevolent advisor genuinely cares about the best interests of the advisee, has a positive attitude towards the advisee, and thinks about the advisee's interests at least as much as her owns.

Even when the advisor has no explicit vested interest in the situation, benevolence can contribute to trustworthiness independently of competence. For exemple, the mere fact that the advisee already knows the advisor (a proxy for benevolence) makes a difference to the advisor's perceived trustworthiness, even when controlling for the advisor's expressed confidence in her advice (a proxy for competence); in fact, this expressed confidence no longer affects trustworthiness as soon as the advisor and the advisee know each other [32]. In these experiments, an increase in trustworthiness translated into a greater weight put on the advisor's recommendation. Other experimental studies directly made it clear to decision makers whether or not

some advisor was benevolent, concerned about their best interests. These experiments concured that recommendations from benevolent advisors are given greater weight in the decision [33].

Interestingly, it has been claimed that people are ready to trade off competence for benevolence when the emotional load of their decision is high [33]. One experiment put subjects in a situation to decide whether they would leave their savings in a badly performing fund, or take them out. In the low emotional load condition, the savings were meant to pay for a summer band camp for young musicians. In that case, subjects sought competent rather than benevolent advisors. In the high emotional load condition, the savings were meant to pay for college. In that case, subjects sought benevolent advisors, and were ready to sacrifice some level of competence in order to ensure benevolence.

Whether this effect is truly due to emotional load or to another confounded variable is not quite clear, but the possibility of a trade off between competence and benevolence would already be especially relevant to our current purpose, given that we conceptualise competence and benevolence as different dimensions of the complex concept of trust. It would mean that a global index of trust might not be precise enough to accomodate people's needs. Indeed, different situations may require different mix of competence and benevolence, although the global index of trust would remain the same.

Benevolence-based trust can obviously be harmed by malevolent behavior. However, it can be repaired on the long term by subsequent benevolent behavior, or, on the short term, by promises to adopt a benevolent behavior. Apologies for malevolent behavior do not seem sufficient, though, to repair trust [28].

3.3 Integrity

The integrity of the advisor reflects her unconditional adherence to a set of principles deemed appropriate by the advisee. Note that integrity so defined can be independent of benevolence. For example, one may expect an advisor to maintain confidentiality whether or not one believes the advisor to be benevolent. Conversely, one may question whether an advisor can be trusted to maintain confidentiality, independently of whether this advisor is benevolent or not.

Some indices of trust put a strong emphasis of integrity. For example, recent studies investigating the relation between emotion, trust, and the uptake of advice [7, 10] used a measure of trust that focused on whether the advisor could be expected to unconditionally honor commitments, and whether the advisor could be expected to unconditionally tell the truth. These studies

found that incidental emotions (i.e., which were felt independently of the advisor) could affect this integrity-based trust, which affected in turn the weight given to the advisor's recommendation. More specifically, incidental anger decreased integrity-based trust, and incidental gratitude or happiness increased integrity-based trust.

Finally, integrity-based trust seems hard to repair once it has been harmed by a failure to honor one's commitment [28]. Once an individual has failed to deliver on a promise, her trustworthiness appears to be durably impaired, and not significantly repaired by apologies or renewed promises to change behavior, even when these promises are genuinely honored.

3.4 Summary

Agents faced with difficult decisions often find that they do not possess all the knowledge and expertise required to make the best possible choice. A natural solution is then to seek expert recommendation about the decision; but because experts may have vested interests in the consequences of their recommendation, they need to be trusted by the agent making the decision. Our global world offers easy access to a vast pool of experts; but it does not offer the traditional guarantees of trustworthiness that come with a history of personal interaction with all these experts.

This problem of computer-mediated trust in expert recommendations clearly falls within the scope of the distributed cognition framework proposed by Hollan and collaborators [14]. Indeed, it presents the three following characteristic features:

- Cognitive processes are distributed accross the members of the social group. Not only is the final decision codependent on computations made by the decision maker and by the expert, but the trust granted to the expert is itself the result of distributed computation among the users of the platform.

- Cognitive processes involve coordination between internal and external structure. To reach an overall assessment of trustworthiness, the decision maker cannot simply inquire into the judgments made by others, but must delegate some computations to the platform and coordinate with the results of this computation.

- Processes are distributed through time in such a way that the products of earlier events transform the nature of later events. Indeed, the dynamics of trust is such that events cannot be interpreted in isolation. A display of integrity, for example, has a very different impact

on trustworthiness depending on whether the expert is known to have given at least one dishonest recommendation.

Overall, the computer-mediated construction of trust is a distributed cognitive process exhibiting a complex trajectory over agents, events, and time, and requires coordination with an external computational structure. It does not result, however, in any radical conceptual rewiring of the nature of mind or trust. In that sense, we offer a 'weak' distributed perspective, focused on the multi-level aggregation of the cognitive outputs of humans and artefacts: a formally difficult problem, but a tractable one.

Our approach sticks to a conceptualisation of the mind as an information processing system, with clearly defined inputs and outputs; and our work rests on the assumption that a significant portion (though clearly not the whole) of trust-building boils down to information processing. Although some aspects of trust-building elude our formalization, we believe that the cold information processes captured by our formalization can already offer some solid decisional grounds. These processes are constrained by a number of variables and psychological dimensions, which we explored in the previous section. In line with previous psychological research, we conceptualise trust as a multidimensional concept comprising competence (expert ability), benevolence (positive attitude and concern towards the interests of the advisee), and integrity (unconditional adherence to a set of principles deemed appropriate by the advisee).

These three dimensions of trust exhibit different degrees of asymmetry in the differential impact of positive and negative information. In the case of integrity, negative information receives extremely greater weight than positive information. This asymmetry is also observed with respect to competence, but apparently to a lesser extent. Finally, the asymmetry would appear to be the least pronounced in the case of benevolence.

Some compensation seems possible between the dimensions of competence and benevolence, since situations appear to exist where advisees are willing to sacrifice some mesure of competence to ensure some mesure of benevolence. It is less clear whether integrity can be traded that way, or whether it should be considered as a completely non-compensatory dimension of trust. One possibility, that would need empirical validation, is that the level of integrity functions as the upper-bound for the level of trustworthiness. A related solution to the problem of computer-mediated trust is to first filter out advisors who have been judged to lack integrity; and then to provide the user with aggregated indices of competence and benevolence, without taking the responsibility to trade one for another in a global index of trustworthiness. This responsibility should be left to the user, who knows best whether

the situation primarily calls for competence, benevolence, or both.

We now turn to the formal treatment of our problem. We introduce a logical framework wherein the three aspects of trust can be formally characterized, and wherein we can model trust reasoning about these three aspects.

4 Formal Treatment

This section presents a logical framework called $\mathcal{TRUST}$ in which the competence, benevolence and integrity of an advisor can be formally characterized. $\mathcal{TRUST}$ is a multi-modal logic which supports reasoning about time, agents' actions and agents' mental attitudes including beliefs and goals. It also allows to express the normative concept of obligation. In this sense, $\mathcal{TRUST}$ combines the expressiveness of dynamic logic [11], temporal logic [8] and deontic logic [1] with the expressiveness of a so-called BDI (belief, desire, intention) logic of agents' mental attitudes [5]. We introduced the logic $\mathcal{TRUST}$ in our previous works on the logical formalization of the concepts of trust and reputation [19]. It is not the aim of this work to discuss the precise semantics of the modal operators of the logic $\mathcal{TRUST}$. We just present them in an informal way by highlighting their intuitive meanings and their basic properties.[1]

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

- a nonempty finite set of agents $AGT = \{i, j, \ldots\}$;
- a nonempty finite set of atomic actions $AT = \{a, b, \ldots\}$;
- a finite set of propositional atoms $ATM = \{p, q, \ldots\}$.

The language of $\mathcal{TRUST}$ is defined as the smallest superset of ATM such that:

- if $\varphi, \psi \in \mathcal{L}$, $\alpha \in ACT$ and $i \in AGT$ then
  $\neg\varphi, \varphi \vee \psi, \mathtt{Does}_{i:\alpha}\,\varphi, \mathtt{Bel}_i\,\varphi, \mathtt{Choice}_i\,\varphi, \mathtt{Past}\varphi, \mathtt{Obl}\varphi \in \mathcal{L}$.

ACT is the set of complex actions and is defined as follows:

$$ACT = AT \cup \{inf_j(\varphi) | j \in AGT,\ \varphi \in \mathcal{L}\}.$$

An action of the form $inf_j(\varphi)$ denotes the action of informing agent j that φ is true. We call this kind of actions informative actions.

[1] See for instance [19] for an analysis of the semantics of these operators, their relationships, and their correspondence with the structural conditions on the models of the logic $\mathcal{TRUST}$.

Thus, the logic $\mathcal{TRUST}$ has five types of modalities: $\mathtt{Bel}_i$, $\mathtt{Choice}_i$, $\mathtt{Does}_{i:\alpha}$, $\mathtt{Past}\varphi$ and $\mathtt{Obl}$. These modalities have the following intuitive meaning.

- $\mathtt{Bel}_i\,\varphi$: the agent i believes that φ;
- $\mathtt{Does}_{i:\alpha}\,\varphi$: agent i is going to do α and φ will be true afterward ($\mathtt{Does}_{i:\alpha}\,\top$ is read: agent i is going to do α);
- $\mathtt{Past}\varphi$: it has at some time been the case that φ;
- $\mathtt{Choice}_i\,\varphi$: agent i has the chosen goal that φ holds (or simply agent i wants that φ holds).

Operators of the form $\mathtt{Choice}_i$ are used to denote an agent's chosen goals, that is, the goals that the agent has decided to pursue. We do not consider how an agent's chosen goals originate through deliberation from more primitive motivational attitudes called desires (see e.g. [25, 6, 3] on this issue).

The following abbreviations will be convenient:

$$\begin{aligned} \mathtt{Intends}_i(\alpha) &\stackrel{\text{def}}{=} \mathtt{Choice}_i\,\mathtt{Does}_{i:\alpha}\,\top \\ \mathtt{Inf}_{i,j}(\varphi) &\stackrel{\text{def}}{=} \mathtt{Does}_{i:inf_j(\varphi)}\,\top \\ \mathtt{BelIf}_i\varphi &\stackrel{\text{def}}{=} \mathtt{Bel}_i\,\varphi \vee \mathtt{Bel}_i\,\neg\varphi \end{aligned}$$

$\mathtt{Intends}_i(\alpha)$ stands for 'agent i intends to do action α'. This means that i's intention to perform action α is defined by agent i's choice to perform action α. $\mathtt{Inf}_{i,j}(\varphi)$ stands for 'agent i informs agent j that the fact φ is true'. Finally, $\mathtt{BelIf}_i\varphi$ stands for 'agent i believes whether φ is true'.

Operators for actions of type $\mathtt{Does}_{i:\alpha}$ are normal modal operators satisfying the axioms and rules of inference of the basic normal modal logic K [4].

Operators of type $\mathtt{Bel}_i\,\varphi$ are just standard doxastic operators in Hintikka style [13] satisfying the axioms and rules of inference of the so-called system KD45 [4]. It follows that an agent cannot have inconsistent beliefs, and an agent has positive and negative introspection over his beliefs. Formally:

$\mathbf{D}_{Bel}$ $\quad \neg(\mathtt{Bel}_i\,\varphi \wedge \mathtt{Bel}_i\,\neg\varphi)$

$\mathbf{4}_{Bel}$ $\quad \mathtt{Bel}_i\,\varphi \rightarrow \mathtt{Bel}_i\,\mathtt{Bel}_i\,\varphi$

$\mathbf{5}_{Bel}$ $\quad \neg\mathtt{Bel}_i\,\varphi \rightarrow \mathtt{Bel}_i\,\neg\mathtt{Bel}_i\,\varphi$

As emphasized above, operators of the form $\mathtt{Choice}_i$ express an agent's chosen goals. These are similar to the modal operators studied in [5]. 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 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). Thus, every operator $\texttt{Choice}_i$ is supposed to satisfy the axioms and rules of inference of the so-called system KD [4]. It follows that an agent cannot choose φ and $\neg\varphi$ at the same time. Moreover chosen goals have to be compatible with beliefs. Formally:

$\mathbf{D}_{Choice}$ $\quad \neg(\texttt{Choice}_i\,\varphi \wedge \texttt{Choice}_i\,\neg\varphi)$

$\mathbf{Comp}_{Bel,Choice}$ $\quad \texttt{Bel}_i\,\varphi \rightarrow \neg\texttt{Choice}_i\,\neg\varphi$

As far as the modal operator for obligation is concerned, we take the operator of Standard Deontic Logic (SDL) [1]. That is, the modality $\texttt{Obl}$ is also supposed to satisfy the axioms and rules of inference of the so-called system KD. It follows that obligations have to be consistent. That is:

$\mathbf{D}_{Obl}$ $\quad \neg(\texttt{Obl}\varphi \wedge \texttt{Obl}\neg\varphi)$

The temporal operator $\texttt{Past}\varphi$ is also a normal modality which satisfies the axioms and rules of inference of system of the basic normal modal logic K. The following two additional axioms are added in order to capture two essential aspects of time.

$\mathbf{4}_{Past}$ $\quad \texttt{PastPast}\varphi \rightarrow \texttt{Past}\varphi$

$\mathbf{Connected}_{Past}$ $\quad (\texttt{Past}\varphi \wedge \texttt{Past}\psi) \rightarrow$
$(\texttt{Past}(\varphi \wedge \texttt{Past}\psi) \vee \texttt{Past}(\psi \wedge \texttt{Past}\varphi) \vee \texttt{Past}(\psi \wedge \varphi))$

Axiom $\mathbf{4}_{Past}$ says that the past satisfies transitivity: if it has been the case that it has been the case that φ then it has been the case that φ. Axiom $\mathbf{Connected}_{Past}$ just expresses that the past is connected: if there are two past moments t and t' then either t is in the past of t' or t' is in the past t or $t = t'$.

Other relationships between the different modalities of the logic $\mathcal{TRUST}$ are expressed by the following logical axioms.

$\mathbf{Alt}_{Does}$ $\quad \texttt{Does}_{i:\alpha}\,\varphi \rightarrow \neg\texttt{Does}_{j:\beta}\,\neg\varphi$

$\mathbf{IntAct}$ $\quad \texttt{Does}_{i:\alpha}\,\top \rightarrow \texttt{Intends}_i(\alpha)$

$\mathbf{Inc}_{Time,Does}$ $\quad (\varphi \wedge \texttt{Does}_{i:\alpha}\,\top) \rightarrow \texttt{Does}_{i:\alpha}\,\texttt{Past}\varphi$

Axiom $\mathbf{Alt}_{Does}$ says that: if agent i is going to do α and φ will be true afterward, then it cannot be the case that agent j is going to do β and $\neg\varphi$ will be true afterward. Axiom **IntAct** relates an agent's intentions with his actions. According to this axiom, an agent is going to do action α only if has the intention to perform action α. In this sense it is supposed that an agent's *doing* is by definition intentional. A similar axiom has been studied in [20, 21] in which a logical model of the relationships between intention and action performance is proposed. Finally Axiom $\mathbf{Inc}_{Time,Does}$ expresses that every action occurrence goes from the present to the future (i.e. actions do not go back to the past). That is, if φ is true in the present and agent i does action α then, after the occurrence of action α, φ is true at some point in the past.

4.1 Formal definitions of competence, benevolence and integrity

The aim of this section is to formalize in the logic $\mathcal{TRUST}$ the three properties competence, benevolence and integrity of a potential advisor.

We start with the concept of competence of an advisor to provide good recommendations about a certain issue φ.

Definition 1 ***Competence****. Agent j is a competent advisor (or competent information source) about a certain issue φ if and only if, if agent j believes that φ then φ is true.*

This notion of competence can be formally expressed as follows:

$$\texttt{Competent}_j(\varphi) \stackrel{\text{def}}{=} \texttt{Bel}_j\,\varphi \rightarrow \varphi.$$

The second concept we aim at formalizing is benevolence.

Definition 2 ***Benevolence****. Agent j is a benevolent advisor (or benevolent information source) about a certain issue φ if and only if, for every agent i, if j believes that i wants to believe whether φ is true and j believes that φ is true then j informs i about his opinion.*

This notion of benevolence can be formally expressed as follows:

$$\texttt{Benevolent}_j(\varphi) \stackrel{\text{def}}{=} \bigwedge_{i\in AGT}((\texttt{Bel}_j\,\texttt{Choice}_i\,\texttt{BelIf}_i\varphi \wedge \texttt{Bel}_j\,\varphi) \rightarrow \texttt{Inf}_{j,i}(\varphi)).$$

As far as integrity is concerned, we split this concept into three different concepts of sincerity, confidentiality and obedience. That is, we suppose that the expression 'the advisor satisfies the property of integrity' means that the advisor is sincere, obedient, and he guarantees the confidentiality of the information.

Definition 3 ***Sincerity****. Agent j is a sincere advisor (or sincere information source) about a certain issue φ if and only if, for every agent i, if j informs i that φ is true then j believes that φ is true.*

This notion of sincerity can be formally expressed as follows:

$$\texttt{Sincere}_j(\varphi) \stackrel{\text{def}}{=} \bigwedge_{i \in AGT}(\texttt{Inf}_{j,i}(\varphi) \rightarrow \texttt{Bel}_j\,\varphi).$$

Definition 4 ***Confidentiality (or Privacy)****. Agent j is an advisor (or information source) which guarantees the confidentiality (or privacy) of the information φ if and only if, for every agent i, if it is obligatory that j does not inform i that φ is true then j does not inform i that φ is true.*

This notion of confidentiality can be formally expressed as follows:

$$\texttt{Privacy}_j(\varphi) \stackrel{\text{def}}{=} \bigwedge_{i \in AGT}(\texttt{Obl}\neg\texttt{Inf}_{j,i}(\varphi) \rightarrow \neg\texttt{Inf}_{j,i}(\varphi)).$$

Definition 5 ***Obedience****. Agent j is an obedient advisor (or obedient information source) about a certain issue φ if and only if, for every agent i, if j is obliged to inform i about φ then j informs i about φ.*

This notion of obedience can be formally expressed as follows:

$$\texttt{Obedient}_j(\varphi) \stackrel{\text{def}}{=} \bigwedge_{i \in AGT}(\texttt{Obl Inf}_{j,i}(\varphi) \rightarrow \texttt{Inf}_{j,i}(\varphi)).$$

We define the integrity of the advisor j about a certain issue φ as the logical conjunction of j's sincerity about φ, j's obedience about φ, the fact that j guarantees the confidentiality of the information φ:

$$\texttt{Integrity}_j(\varphi) \stackrel{\text{def}}{=} \texttt{Sincere}_j(\varphi) \wedge \texttt{Privacy}_j(\varphi) \wedge \texttt{Obedient}_j(\varphi).$$

4.2 Trust reasoning about competence, benevolence and integrity

When assessing the trustworthiness of a certain advisor k, the truster i evaluates whether k has the three properties of competence, benevolence and integrity. In many situations, such an evaluation might depend on what agent i has heard about the advisor k in the past. In particular, agent i's evaluation of an agent k's competence, benevolence and integrity might be based on what the other agents told to i about k. In these situations, agent i has to apply certain procedures for *aggregating* all information that he has received from the other agents about k's properties.

The logic $\mathcal{TRUST}$ allows to formalize some of these procedures, namely *majority* and *unanimity*. For instance, we can specify the concept of 'the majority of agents informed agent i that agent k is benevolent about φ'.

$$\texttt{Maj}_i(\texttt{Benevolent}_k(\varphi)) \stackrel{\text{def}}{=} \bigvee_{J \subseteq AGT, |J|>|AGT \setminus J|} (\bigwedge_{j \in J} \texttt{Past Inf}_{j,i}(\texttt{Benevolent}_k(\varphi))$$

According to this definition, the majority of agents informed agent i that agent k is benevolent about φ (noted $\texttt{Maj}_i(\texttt{Benevolent}_k(\varphi))$) if and only if there exists a group of agents J such that every agent j in J informed i that k is benevolent about φ and J is larger than its complement with respect to AGT.

In a similar way we can express that 'the majority of agents informed agent i that agent k is competent about φ'.

$$\texttt{Maj}_i(\texttt{Competent}_k(\varphi)) \stackrel{\text{def}}{=} \bigvee_{J \subseteq AGT, |J|>|AGT \setminus J|} (\bigwedge_{j \in J} \texttt{Past Inf}_{j,i}(\texttt{Competent}_k(\varphi))$$

As far as unanimity is concerned, we can specify the concept of 'all agents unanimously informed agent i that agent k satisfies the property of integrity'.

$$\texttt{Unan}_i(\texttt{Integrity}_k(\varphi)) \stackrel{\text{def}}{=} \bigwedge_{j \in AGT} \texttt{Past Inf}_{j,i}(\texttt{Integrity}_k(\varphi))$$

The previous definitions of majority-based benevolence and competence and unanimity-based integrity can be used to specify the procedures adopted by agent i to evaluate a certain advisor k. From the experimental literature that we reviewed in Section 3, it seems sensible to use a strong unanimity procedure for integrity, but to allow a more lenient majority procedure for competence and benevolence:

$$\texttt{Maj}_i(\texttt{Competent}_k(\varphi)) \rightarrow \texttt{Bel}_i\, \texttt{Competent}_k(\varphi).$$

This rule says that if the majority of agents informed i that k is a competent advisor then i believes so.

$$\texttt{Maj}_i(\texttt{Benevolent}_k(\varphi)) \rightarrow \texttt{Bel}_i\, \texttt{Benevolent}_k(\varphi).$$

This rule says that if the majority of agents informed i that k is a benevolent advisor then i believes so.

$$\texttt{Unan}_i(\texttt{Integrity}_k(\varphi)) \rightarrow \texttt{Bel}_i\, \texttt{Integrity}_k(\varphi).$$

This rule just says that if all agents informed i that k is an advisor which satisfies the property of integrity then i believes so.

At this point, and although much has still to be articulated, we will conclude the formal analysis of our problem. Indeed, our goal in this article has not been to solve the problem of computer-mediated trust in partial expert recommendations, but rather to provide a roadmap for addressing this problem, by integrating findings from experimental psychology and formal tools from Artificial Intelligence. In the last section of this article, we go beyond our initial problem by suggesting extended applications of our approach, to a range of problems where trust (or reputation) cannot be assessed by personal interaction, where agents cannot be vouched for by an objective arbiter, but where the possibility remains of applying some variant of our approach.

5 Extended Applications

In its most general formulation, the problem we have tackled here concerns multi-agent applications where users have to evaluate (or simply compare) agents, but it is impossible to call on an objective arbiter to provide some help. This may happen for various reasons, for example, the number of agents is too large, no arbiter is considered sufficiently competent and sincere, arbiters are too expensive, etc. However, in such applications, a lot of feedbacks may be available, that is, information about agents provided by other agents. Trust and reputation systems of the kind we have envisioned here are conceived to exploit such information in order to help users to take decisions about other agents.

The information provided by peers should be used with caution. It can be incomplete, and it may be downright false. Indeed, agents may have vested interests in their judgments, and therefore may lie or hide the truth to serve their interests. Another issue that is critical in any trust and reputation system is that of *cycles* of information (e.g., a provides information about b, b provides information c, and c provides information a). Trust and reputation systems have to give different weights to the pieces of information provided by the agents, but assigning such weights in a rational way turns out to be difficult in the presence of information cycles. In this final section, we consider several situations where a trust and reputation system can be used to overcome the absence of a neutral, objective arbiter.

Currently, the best-known examples of a virtual community of agents are social networks such as of Facebook or MySpace. We briefly evoke this setting because of its popularity, alghough it does not, strictly speaking, relate to our topic; indeed, the reputation system that can be implemented in this setting is likely to be gratuitous (it is not meant as a decision help) and unrelated to our central issue of trust. Still, in such a social network, a wealth of information is given by agents about other agents. For example, in addition to the comments and pictures they leave on each other 'walls', users can rate their virtual friends on a number of dimensions (are they attractive? honest? serious?), or vote for the nicest person in their network; and all this information can be used to extract aggregated judgments about any particular user.

Other applications are, to a greater extent, geared to help decisions. For example, e-commerce applications like Ebay are such systems where it is useful to have information about sellers before deciding to buy an item. Here, the agents are the users and the dimension of trust that is the most decisive is integrity, the expectation that the seller respects his commitments and tell the truth. In this kind of system, there are too many buyers and sellers for an

external arbiter to evaluate them all. However, after each transaction, buyer and seller have an opportunity to appraise each other. This rich amount of feedback can be exploited to reach aggregated evaluation of individual ebayers. Ebay is already equipped with a simple reputation system, which does not however explicitly measure a score of integrity-based trust. Rather, it uses a simple scheme where a positive feedback from a buyer brings one point, a negative feedback removes one point, and a neutral feedback has no consequences. Symbolic trinkets are attached to some scores (e.g., a star when the seller reaches a net score of 10 points). One limitation of this system is that it does not weight feedback according to the reputation of the ebayer who provided it.

Agents in a trust and reputation system need not be human. Indeed, web pages may be seen as agents, and a link between two pages may be construed as a positive recommendation by the linking page about the linked page. Pages that gathered the most agreggated support can be considered as more trustworthy along the competence dimension of trust: they are the pages were relevant information is to be found. This is in fact one of the broad principles that PageRank (the reputation system used by the Google search engine) is based on.

Scientific citation indices offer another application of trust and reputation systems, where scientific papers are the agents (or, perhaps, the minions of the scientists who wrote them). In most scientific reputation indices, citing a paper is construed as a positive recommendation of that paper. This is true of very basic indices such as the raw number of citations, as well as of more elaborated indices such as the h index. As often, this framework gives every citation the same weight in the aggregated evaluation of the paper or the scientist who wrote the collection of papers under consideration. A trust and reputation system would allow to weight a citation according to aggregated scores of the citing paper that would take into account the potential for vested interests in citing one article rather than another.

The last application we consider in this discussion is less publicized, partly because of its more technical nature. It concerns the important issue of message encryption and public key certificates. Without engaging in too technical a discussion, we can summarize the problem as follows. Various agents wish to exchange encrypted messages. Each agent is in possession of two personal keys. One of these is public, it can be used to encrypt messages sent to this agent; the other is private, it is used by the agent to decrypt messages that were encrypted using his or her public key.

One concern within this framework is that a malicious agent may assume the identity of another agent a, and pretend that her own public key is actually the public key of a. Other agents may then mistakenly believe they

are encrypting messages with the public key of a, when they are really using the public key of the malicious agent. The malicious agent can then intercept and decrypt messages that were meant for a.

The problem for any agent, then, is to have sufficient ground to believe that what is advertized as the public key of a truly is the public key of a. Public key certificates are used to solve that problem. A public key certificate is a document supposedly written by an agent b, signed with a public key K_b, that certifies that an agent a is the owner of a public key K_a. Consequently, we can extract from this framework a set of pairs composed of an agent and a public key supposedly belonging to it. In addition, we can extract a binary support relation between these pairs. More precisely, we consider that a pair (b, K_b) supports the credibility of a pair (a, K_a) if there exists a certificate from b, signed with K_b, stating that a is the owner of K_a. This information can be used to evaluated the credibility of the different pairs. For example, the well-known Pretty Good Privacy system looks for chains of pairs, where the credibility of the first element is trusted by the sender, each element supports the next one, and the last element is the receiver.

The main limitation of this framework is its extreme cautiousness. If the chain of certification does not go back to some agent trusted by the potential sender, no encrypted message can be sent. One way to overcome this extreme cautiousness (at some risk), is to use the kind of trust and reputation system that we have considered through this article, and to appraise the trustworthiness of an agent-key pair based on the structure of the certification graph.

We do not consider this application in any greater detail, for our goal in this last section was rather to give a broad perspective of the various problems that can be tackled by the general approach we have outlined in this paper. We hope that the reader will have gained a sense of the many domains where a trust and reputation system can help appraise the characteristics of some agents who cannot be evaluated by a central, neutral authority. These applications must be supported by a mix of psychological findings and artificial intelligence formalisms, whose exact composition depends on the extent to which the agents in the system are human or human-like in their behavior and intentions.

References

[1] L. Åqvist. Deontic logic. In D. M. Gabbay and F. Geunther, editors, *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht, 2002.

[2] B. Barber. *The logic and limits of trust*. NJ: Rutgers Univ. Press, 1983.

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

[4] B. F. Chellas. *Modal logic: an introduction*. Cambridge University Press, Cambridge, 1980.

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

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

[7] J. R. Dunn and M. E. Schweitzer. Feeling and believing: the influence of emotion on trust. *Journal of Personality and Social Psychology*, 88:736–748, 2005.

[8] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics*. North-Holland Pub. Co./MIT Press, 1990.

[9] D. L. Ferrin, K. T. Dirks, and P. P. Shah. Direct and indirect effects of third-party relationships on interpersonal trust. *Journal of Applied Psychology*, 91:870–883, 2006.

[10] F. Gino and M. E. Schweitzer. Blinded by anger or feeling the love: How emotions influence advice taking. *Journal of Applied Psychology*, 93:1165–1173, 2008.

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

[12] N. Harvey and I. Fischer. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, 70:117–133, 1997.

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

[14] J. Hollan, E. Hutchins, and D. Kirsh. Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7:174–196, 2000.

[15] T. A. Ito and J. T. Cacioppo. Variations on a human universal: Individual differences in positivity offset and negativity bias. *Cognition and Emotion*, 19:1–26, 2005.

[16] C. E. Johnson-George and W. C. Swap. Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other. *Journal of Personality and Social Psychology*, 43:1306–1317, 1982.

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

[18] J. D. Lewis and A. Weigert. Trust as social reality. *Social Forces*, 63:967–985, 1985.

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

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

[21] E. Lorini, A. Herzig, and C. Castelfranchi. Introducing "attempt" in a modal logic of intentional action. In *Logics in Artificial Intelligence: 10th European Conference (JELIA 2006)*, volume 4160 of *LNAI*, pages 280–292. Springer, 2006.

[22] W. W. Maddux, E. Mullen, and A. D. Galinsky. Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negociation outcomes. *Journal of Experimental Social Psychology*, 44:461–468, 2008.

[23] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20:709–734, 1995.

[24] D. J. McAllister. Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal*, 38:24–59, 1995.

[25] 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.

[26] J. K. Rempel, J. G. Holmes, and M. D. Zanna. Trust in close relationships. *Journal of Personality and Social Psychology*, 49:95–112, 1985.

[27] M. Rousseau, S. Sitkin, R. Burt, and C. Camerer. Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 23:393–404, 1998.

[28] M. E. Schweitzer, J. C. Hershey, and E. T. Bradlow. Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, 101:1–19, 2006.

[29] J. J. Skowronski and D. E. Carlston. Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, 105:131–142, 1989.

[30] P. Slovic. Perceived risk, trust and democracy. *Risk Analysis*, 13:675–685, 1993.

[31] J. A. Sniezek and B. Buckley. Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, 62:159–174, 1995.

[32] L. M. Van Swol and J. A. Sniezek. Trust, confidence and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, 84:288–307, 2001.

[33] T. B. White. Consumer trust and advice acceptance: The moderating roles of benevolence, expertise, and negative emotions. *Journal of Consumer Psychology*, 15:141–148, 2005.

[34] I. Yaniv and E. Kleinberger. Advice taking in decision making: egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83:260–281, 2000.

[35] I. Yaniv, J. F. Yates, and J. E. K. Smith. Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, 110:611–617, 1991.