

# Computational Opinions

Felix Fischer<sup>1</sup> and Matthias Nickles<sup>2</sup>

**Abstract.** Existing approaches to knowledge representation and reasoning in the context of open systems either deal with “objective” knowledge or with beliefs. In contrast, there has been almost no research on the formal modelling of *opinions*, i.e., communicatively asserted ostensible beliefs. This is highly surprising, since opinions are in fact the only publicly visible kind of knowledge in open systems, and can neither be reduced to objective knowledge nor to beliefs. In this paper, we propose a formal framework for the representation of dynamic, context-dependent and revisable *opinions* and *ostensible intentions* as a sound basis for the external description of agents as obtained from observable communication processes. Potential applications include a natural semantics of communicative acts exchanged between truly autonomous agents, and a fine-grained, statement-level concept of trust.

## 1 INTRODUCTION

In the area of artificial intelligence, open systems with heterogeneous, more or less inscrutable actors, are becoming an increasingly important area of application for multi-agency. Prominent examples include the Semantic Web or Peer-to-Peer systems. Nevertheless, agents (including human users) are still mainly characterised in terms of their mental attitudes. The problem with this approach is that agents’ internal states are, by definition, concealed from the point of view of external observers, including other agents. And even if the mental states of all participants were accessible, it would be an extremely difficult task to describe the dynamic behaviour of a system as a function of all of its individuals, particularly if the number of actors is large (as it is usually the case in Web and Peer-to-Peer applications). Hence, although the description of an agent in terms of his beliefs, intentions and desires is a very natural and powerful tool, it reaches its limits rather quickly in complex open environments.

One way to cope with this situation, in order to ensure the controllability of individual agents as well as the achievement of supra-individual system goals, is to impose normative restrictions on agent behaviour, or to model the entire system in terms of static organisational structures. While such top-down approaches certainly have their merits, and systems without predefined policies and protocols are hardly imaginable, they may seriously limit the desirable properties of autonomous systems, like their robustness and flexibility. For example, the use of an agent communication language with a predefined mentalistic semantics [see, e.g., 4] may reduce the level of agent autonomy by prescribing to some degree what the communication parties have to “think” during communication. This is not to dismiss social structures and norms, of course. Public communication models like an ACL semantics cannot avoid normativity if they are to ensure that meaningful communication can take place. We

think, however, that normativity should be kept to a minimum, and the focus should be on adaptive models and on the measurement of the degree to which agents adhere to given models at run-time. After all, agent technology is basically a bottom-up approach, allowing for the emergence of behaviour and solutions, and should not be turned upside down.

Another approach, somewhat complementary to the ones just described, is to constrain the modelling of black-box agents from an external point of view by limiting the *validity* of cognitive models or commitments, thus inducing some sort of bounded observer rationality. A particular way of doing this is by using information about the trustability and reputation of the respective agent. The approach introduced in [13] and refined in this paper is largely compatible and in line with such means, but takes a different perspective. Instead of adding restrictions to statements about internal agent properties and commitments, we introduce the *communication attitudes* (CAs) *opinion* (also called ostensible belief) and *ostensible intention* as communication-level counterparts to belief and intention, and thus “lift” mental attitudes to the social level. As for the intended effect, this is similar to approaches using *social commitments* [18]. However, the latter have not yet been fully formalised, and there currently exists no consensus about the precise meaning of “being committed”. In contrast, CAs will be given an intuitive but precise formal semantics which resembles the modalities of agent belief and intention in many ways, and can be used as a direct replacement for belief and intention in most imaginable scenarios (e.g., in the semantics of KQML/KIF or FIPA ACL). CAs are nevertheless cleanly separated from mental attitudes, and can thus be used together with these without any interference. For example, an agent might reason simultaneously about the (alleged) “real” beliefs of another agent and the opinions held communicatively by this agent.

To the best of our knowledge, the formal, systematic treatment of opinions in the above sense, i.e., by distinguishing between rational mental attitudes on the one hand and (boundedly) rational yet “superficial” stances an agent exposes in discourse on the other, is new in the field of AI. This is highly surprising since opinions are in fact the only publicly visible kind of knowledge in open systems, and can neither be reduced to objective knowledge nor to beliefs. Unlike objective knowledge, opinions need not be grounded, shared or consistent. What distinguishes an agent’s CA of opinion from his mental attitude of belief is that the former is triggered and revised by social conditions, more precisely communicative acts and underlying social structures like legal, organisational and economic laws. While opinions might as well reflect the true beliefs of benevolent, trustworthy agents (an assumption that is made by most traditional approaches to agent communication semantics), this should be considered a special case for autonomous, self-interested agents in open systems. Also, the utterance of an opinion should not mean that one truthfully intends to make someone adopt this opinion as a belief (which would be unrealistic), but rather as an opinion. Opinions emerge from (more or less hidden) agent intentions and social processes, they are tailored to the intended communicative effect and

<sup>1</sup> Computer Science Department, University of Munich, 80538 Munich, Germany, email: fischerf@tcs.ifi.lmu.de

<sup>2</sup> Computer Science Department, Technical University of Munich, 85748 Garching, Germany, email: nickles@in.tum.de

to the opinions and ostensible intentions of the audience. Thus, traditional concepts like epistemic logic, knowledge acquisition, belief ascription and revision, etc., do not necessarily apply to opinions (for example, an opinion could be bought in exchange for money). Yet another approach, which many people use at least implicitly, would be to “abuse” mental attitudes and combine them with additional assumptions like the trustability of agents to model agent interaction. Related but somewhat more appropriate is the ascription of *weak intentions*, which are not necessarily pursued until the intended effect has been achieved or is deemed impossible [2]. This seems workable, since even ostensible intentions and opinions have to built upon some undeniable mental intentions (totally unconscious communication exempted), possibly with a duration of self-commitment that is shorter than claimed. However, it would again lead to problems with cognition and sociality getting mixed up, and force a socially reasoning communication observer to find an immediate, possibly complicated mapping of alleged attitudes to weak intentions. We think that such an amalgamation of mental and pseudo-communication attitudes, while still common, is dangerous and inappropriate. It is likely to impede the simultaneous reasoning about mental and social issues, and it is misleading with respect to the very nature of communication processes and assertions, which form a system of their own that is in a certain sense decoupled from the underlying mental cognition processes [12]. While we will also introduce a notion of trust tailored to the use with opinions, our approach resides on a different conceptual level than related areas like belief revision and information integration. While the latter are mainly concerned with the determination of correct, consistent and useful information, our primary goal is to represent communicatory properties of and differences between heterogeneous assertions, which precedes a potential assessment in terms of reliability.

## 2 A COMMUNICATION ATTITUDE LOGIC

In this section, we introduce a probabilistic and dynamic *Communication Attitude Logic* CAL as an extension of [7, 2, 1] with additional modalities for opinions and ostensible intentions. Our main goal here is to “lift” the usual modal logic and protocol languages for modelling agent beliefs and intentions and their revision [see, e.g., 8] to the communication (i.e., social) level, in order to deal with mentally opaque agents in open systems. Within the scope of this work, we have chosen a probabilistic variant of dynamic logic as a basis for this, but other formalisms like the *event calculus* [10] could certainly be used instead. Even more importantly, the underlying notions of opinion and ostensible intention do not depend in any way on a dynamic or probabilistic interpretation. As we have pointed out in the previous section, an important property of CAs is their dependency on social conditions, demarcating them, inter alia, from mental attitudes. Since an exhaustive description of all the possible ways by which CAs are steered by social conditions of various kinds is beyond the scope of this paper, we confine ourselves to the presentation of simple trigger events.

### 2.1 Syntax

**Definition 1** *The language  $\mathcal{L}$  of well-formed CAL formulas  $\varphi, \psi$  and processes  $\alpha, \beta$  is given by*

$$\begin{aligned} \varphi, \psi ::= & P(V_1, V_2, \dots) \mid \top \mid \perp \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \\ & \mid \varphi \leftrightarrow \psi \mid \diamond\varphi \mid \square\varphi \mid \langle\alpha\rangle p \mid \text{done}(\alpha) \mid \text{happens}(\alpha) \\ & \mid Op(A_1, A_2, \varphi) \mid OInt(a_1, A_2, \varphi) \mid Bel_d(a_1, \varphi) \mid Int(a_1, \varphi) \end{aligned}$$

$$\alpha, \beta ::= \text{act} \mid a_i.\text{act} \mid \text{any} \mid \alpha;\beta \mid \alpha \cup \beta \mid \alpha^* \mid \varphi?,$$

where

- $a_i \in A$  and  $A_i \subseteq A$  denote an agent or a set of agents, the source or addressee of an opinion;
- $P$  is a predicate symbol,  $V_i$  are variables;
- $\text{act}, a_i.\text{act}$  is an elementary action, possibly indexed with the acting agent  $a_i$ ;
- $\text{any}$  is an arbitrary action;
- $\alpha;\beta$  denotes sequential process combination;
- $\alpha \cup \beta$  denotes non-deterministic choice between  $\alpha$  and  $\beta$ ;
- $\alpha^*$  denotes zero or more iterations of  $\alpha$ ;
- $\varphi?$  is a test action (i.e., the process proceeds if  $\varphi$  holds true);
- $\langle\alpha\rangle p$  denotes that  $\alpha$  is processed and  $p$  holds afterwards;
- $Op(A_1, A_2, \varphi)$  denotes that agents  $A_1$  hold opinion (ostensible belief)  $\varphi$  facing agents  $A_2$ ;
- $OInt(a_1, A_2, \varphi)$  denotes that agent  $a_1$ , facing agents  $A_2$ , exhibits the ostensible intention to make  $\varphi$  become true (either by himself or indirectly via agents  $A_2$ );
- $Bel_d(A_1, \varphi)$  denotes that every agent  $a_1 \in A_1$  (sincerely) believes  $\varphi$  with degree  $d$ ; and
- $Int(a_1, \varphi)$  denotes that  $a_1$  (sincerely) intends  $\varphi$ .

$P(V_1, V_2, \dots), \neg, \wedge, \top, \perp, \rightarrow$  and  $\leftrightarrow$  shall have the usual meaning. The meaning of *done*( $\alpha$ ), *happens*( $\alpha$ ),  $\diamond\varphi$ , and  $\square\varphi$  is that  $\alpha$  has just happened, or is currently taking place, and that  $\varphi$  will hold eventually or always, respectively. The set  $A$  may include both agent and non-agent opinion resources (like a Web document). For the sake of readability, we will henceforth refer to all of these resources as *agents*. We further use the following abbreviations:

$$\begin{aligned} Op(a_1, A_2, \varphi) &\equiv_{\text{def}} Op(\{a_1\}, A_2, \varphi) \\ Op(A_1, a_2, \varphi) &\equiv_{\text{def}} Op(A_1, \{a_2\}) \\ OInt(a_1, a_2, \varphi) &\equiv_{\text{def}} OInt(a_1, \{a_2\}, \varphi) \\ Bel_d(a_1, \varphi) &\equiv_{\text{def}} Bel_d(\{a_1\}, \varphi) \end{aligned}$$

As an example, consider the sequence of negotiation dialogues given in Table 1. An agent  $a_2$  is offering some item  $o$  for sale, and agents  $a_1$  and  $a_3$  are both interested in buying this item. Agents  $a_1$  and  $a_2$  shall be mentally opaque from our point of view, but we shall know the beliefs of agent  $a_3$  (e.g., we could ourselves be  $a_3$ ). First of all, agent  $a_3$  tells agent  $a_2$  that he would be willing to pay \$100 for  $o$  (Step 1a). At the same time, he tells  $a_2$  that  $a_1$  is notoriously unreliable, while privately believing the opposite to be the case (Step 1b). Then,  $a_1$  enters the scene, and  $a_3$  tells  $a_2$  (with  $a_1$  observing) that  $a_1$  is a reliable customer (Step 1c). In a separate dialogue, agent  $a_1$  tells  $a_2$  that he wants to buy item  $o$  for \$150 (Step 2a). Agent  $a_2$  refuses this offer, insincerely telling  $a_1$  that agent  $a_3$  would be willing to pay \$200 for  $o$  (Step 2b).  $a_1$  then asks  $a_3$  whether he is actually willing to pay \$200, as claimed by  $a_2$  (Step 3). This is when he becomes aware that  $a_2$  (seemingly) tried to cheat him, and tries to extort a lower selling price by telling  $a_2$  that he would otherwise reveal the attempted cheat to  $a_1$  (Step 4). This threat is again insincere, as in fact  $a_3$  does not plan to expose  $a_2$  should he refuse the new offer.

The relevant internal states of  $a_1$  and  $a_2$  being invisible, this scenario cannot be modelled in terms of mental attitudes (e.g., within the BDI framework), This not only applies to the obvious cases of Steps 1b and 4, where  $Bel(a_3, \text{reliable}(a_1))$  and  $Op(a_3, a_2, \neg\text{reliable}(a_1))$ , or  $OInt(a_3, a_2, \text{buyFor}50)$  and  $\neg Int(a_3, \text{buyFor}50)$ , hold simultaneously. It is tempting to write  $Int(a_2, Bel(a_1, Int(a_3, \text{buyFor}200)))$  to model the state after Step 2b,

$$\begin{aligned}
&\langle a_3.request(buyFor100, a_2) \rangle OInt(a_3, a_2, buyFor100) & (1a) \\
&Bel(a_3, reliable(a_1)) \wedge \langle a_3.inform(\neg reliable(a_1), a_2) \rangle & (1b) \\
&\quad Op(a_3, a_2, \neg reliable(a_1)) \\
&\langle a_3.inform(reliable(a_1), \{a_2, a_1\}) \rangle Op(a_3, \{a_2, a_1\}, reliable(a_1)) & (1c) \\
&\langle a_1.request(buyFor150, a_2) \rangle OInt(a_1, a_2, buyFor150) & (2a) \\
&\langle a_2.inform(OInt(a_3, a_2, buyFor200), a_1) \rangle & (2b) \\
&\quad Op(a_2, a_1, OInt(a_3, a_2, buyFor200), a_1) \\
&\langle a_1.request(confirm(OInt(a_3, a_2, buyFor200)), a_3) \rangle & (3) \\
&\quad Bel(a_3, Op(a_2, a_1, OInt(a_3, a_2, buyFor200))) \\
&\langle a_3.request(buyFor50, a_2); a_3.inform(done(a_2.refuse)?); & (4) \\
&\quad a_3.inform(\neg OInt(a_3, a_2, buyFor200), a_1), a_2) \rangle \\
&\quad OInt(a_3, a_2, buyFor50) \wedge OInt(a_3, a_2, done(done(a_2.refuse)?); \\
&\quad a_3.inform(\neg OInt(a_3, a_2, buyFor200), a_1))) \\
&\quad \wedge \neg Int(a_3, done(a_3.inform(\neg OInt(a_3, a_2, buyFor200), a_1)))
\end{aligned}$$

**Table 1.** CAL formalisation of a sequence of dialogues. involving speech acts *inform*, *request*, *confirm*, and *refuse*.

but this would be mere speculation. For example,  $a_2$  may just have tried to provoke the reaction of Step 4 and then denounce  $a_3$  as an extortionist. The example further highlights the ability of ostensible attitudes to model mutual inconsistencies between information circulating in a group of agents and in one of its subgroups, where after Steps 1b and 1c we simultaneously have  $Bel(a_3, reliable(a_1))$ ,  $Op(a_3, a_2, \neg reliable(a_1))$ , and  $Op(a_3, \{a_2, a_1\}, reliable(a_1))$ .

## 2.2 Semantics

The model theory of  $\mathcal{L}$  is that of a first-order linear time temporal logic with modal operators for gradual belief and for intention. For belief, we use (a linear time extension of) the probabilistic semantics proposed in [1], which has been shown to include a KD45 modal operator for full belief as a special case. Intention is modelled by a KD modal operator. We further include two modal operators for the CAs of opinion and ostensible intention. Their individual Kripke-style semantics is roughly equivalent to that of beliefs or intentions, a possible axiomatisation will be discussed in more detail in Section 2.3.

**Definition 2** A model is a tuple  $M = (\Theta, E, W, A, (\mu_a)_{a \in A}, (I_a)_{a \in A}, (O_{A_1, A_2})_{A_1, A_2 \subseteq A}, (J_{a_1, a_2})_{a_1 \in A, a_2 \subseteq A}, \Phi)$  where  $\Theta$  is a universe of discourse,  $E$  is a set of primitive event types,  $W \subseteq \{w : [n] \mapsto E \mid n \in \mathbb{N}\}$  is a set of possible courses of events (or worlds) specified as a total function from the non-negative integers up to  $n$  to elements of  $E$ .  $A$  is a set of agents, and for all  $a \in A$ ,  $\mu_a : S \mapsto [0, 1]$  is a discrete probability distribution over situations (i.e., worlds at a particular time step, defined as  $S = \{(w, i) \mid w : [n] \mapsto E, n \in \mathbb{N}, 1 \leq i \leq n\}$ ) for gradual belief, and  $I_a \subseteq S \times W$  is a serial accessibility relation for intentions. For all  $A_1, A_2 \subseteq A$ ,  $a_1 \in A$ ,  $O_{A_1, A_2} \subseteq S \times W$  is a serial, transitive, Euclidean accessibility relation for opinions, and  $J_{a_1, a_2} \subseteq S \times W$  is a serial accessibility relation for ostensible intentions.  $\Phi$  interprets predicates.

Since  $S$  contains at most a denumerable number of situations,  $\mu$  can be naturally extended to a probability distribution over subsets  $T \subseteq S$  by  $\mu(T) = \sum_{s \in T} \mu(s)$ . In particular, we have  $\mu(S) = 1$ .

Formulas of the logical language  $\mathcal{L}$  are interpreted with respect to a tuple  $(M, v, \sigma, i)$ , consisting of a model  $M$ , a variable assignment  $v$  (mapping variables to elements of  $\Theta$ ), a possible world (i.e., possible course of events)  $\sigma$ , and a time step  $i$  in this particular world. The interpretation of formulas is given by the standard interpretation rules

for first-order formulas and for the action-related modal operators *happens* and *done*. We refer to [2] for formal definitions of these rules. The gradual belief operator  $Bel_p$ , the intention operator  $Int$ , the opinion operator  $Op$ , and the ostensible intention operator  $OInt$  are interpreted as follows:

$$\begin{aligned}
(M, v, \sigma, i) \models Bel_p(a, \varphi) & \quad \text{iff } \mu_a(\{s' \mid (M, v, s', i) \models \varphi\}) = p \\
(M, v, \sigma, i) \models Int(a, \varphi) & \quad \text{iff for all } \sigma' \text{ such that} \\
& \quad (\sigma, i, \sigma') \in I_a, (M, v, \sigma', n) \models \varphi \\
(M, v, \sigma, i) \models Op(A_1, A_2, \varphi) & \quad \text{iff for all } \sigma' \text{ such that } (\sigma, i, \sigma') \in \\
& \quad O_{A_1, A_2}, (M, v, \sigma', n) \models \varphi \\
(M, v, \sigma, i) \models OInt(a_1, A_2, \varphi) & \quad \text{iff for all } \sigma' \text{ such that } (\sigma, i, \sigma') \in \\
& \quad J_{a_1, A_2}, (M, v, \sigma', n) \models \varphi
\end{aligned}$$

That is,  $\varphi$  is *gradually believed* by agent  $a$  with *strength*  $p$  iff the measure  $\mu_a$  of all situations where  $\varphi$  holds true is at least  $p$ , and *intended* by  $a$  iff it holds in all situations that are accessible via  $I_a$ .  $\varphi$  is *ostensibly believed* (*ostensibly intended*) by  $A_1$  facing  $A_2$  ( $a_1$  facing  $A_2$ ) iff it holds in all situations that are accessible via  $O_{A_1, A_2}$  ( $J_{a_1, A_2}$ ). While an agent would usually relate the strength of belief to past events (i.e., frequency) in some way, we will not commit to a particular way this is done in the context of this paper (e.g., by giving axioms that relate  $\mu$  to past events).

## 2.3 Communication Attitudes

Even though agents are fully autonomous, CAs should be governed by a set of axioms that separate rational and irrational communicative behaviour. As we have already outlined above, violation of these axioms would, in the most extreme case, disqualify an agent from taking part in (rational) interaction with other agents. The following axioms describe (the consistency of a set of) CAs in a way that is similar to a KD45 logic of belief and a KD logic of intention, but cleanly separated from these mental attitudes:

1.  $Op(A_1, A_2, \varphi \rightarrow \psi) \rightarrow (Op(A_1, A_2, \varphi) \rightarrow Op(A_1, A_2, \psi))$
2.  $Op(A_1, A_2, \varphi) \rightarrow \neg Op(A_1, A_2, \neg \varphi)$
3.  $Op(A_1, A_2, \varphi) \rightarrow Op(A_1, A_2, Op(A_1, A_2, \varphi))$
4.  $\neg Op(A_1, A_2, \varphi) \rightarrow Op(A_1, A_2, \neg Op(A_1, A_2, \varphi))$
5.  $OInt(a_1, A_2, \varphi \rightarrow \psi) \rightarrow (OInt(a_1, A_2, \varphi) \rightarrow OInt(a_1, A_2, \psi))$
6.  $OInt(a_1, A_2, \varphi) \rightarrow \neg OInt(a_1, A_2, \neg \varphi)$
7.  $OInt(a_1, A_2, \varphi) \leftrightarrow Op(a_1, A_2, OInt(a_1, A_2, \varphi))$
8.  $OInt(a_1, A_2, \varphi) \leftrightarrow Op(A_2, a_1, OInt(a_1, A_2, \varphi))$
9.  $\neg OInt(a_1, A_2, \varphi) \leftrightarrow Op(a_1, A_2, \neg OInt(a_1, A_2, \varphi))$
10.  $\neg OInt(a_1, A_2, \varphi) \leftrightarrow Op(A_2, a_1, \neg OInt(a_1, A_2, \varphi))$
11.  $OInt(a_1, A_2, \varphi) \rightarrow \neg Op(a_1, A_2, \varphi)$
12.  $OInt(a_1, A_2, Op(A_2, a_1, \varphi)) \rightarrow Op(a_1, A_2, \varphi)$
13.  $Op(a_1, A_2, \varphi) \rightarrow \langle A_2.query(a_1, \varphi); a_1.reply(A_2, \varphi) \rangle \top$

Axioms 1 to 6 correspond to the usual axioms of a KD45 modal operator of belief and to that of a KD operator of intention, and are additionally contextualised with both sender and addressee. The remaining axioms concern additional properties of opinions, ostensible intentions, and the relationship between them. Axiom 13 states that agents (have to) admit their opinions at least upon request, if this request is observed by the opinion addressees. In practice, the above axioms, which provide a basic, abstract set of social norms, could be augmented by additional rules concerning “etiquette” in social interaction and emergent social norms (e.g., under what circumstances it is socially acceptable to retract an assertion or drop an intention).

For opinions that are not just held and uttered on request (like in an opinion poll), but actively asserted facing other agents, we can

define a special kind of *active opinion*, denoted  $Op^a$ . Since the latter includes the ostensible intention to spread the opinion among the respective addressees, its semantics can directly be given in terms of ostensible intentions, i.e.,

$$Op^a(a_1, A_2, \varphi) \equiv_{def} OInt(a_1, A_2, Op(A_2, a_1, \varphi)).$$

By Axiom 12, we have  $Op^a(a_1, A_2, \varphi) \rightarrow Op(a_1, A_2, \varphi)$ , i.e., *passive opinions* form the consequential closure of those pro-actively put forward in terms of communication. By Axioms 8 and 12, we further have  $Op^a(a_1, A_2, \varphi) \rightarrow Op(A_2, a_1, Op^a(a_1, A_2, \varphi))$ , a variation of Axiom 3 for the addressee. We do not distinguish between active and passive ostensible intentions, because in this case consequential closure mainly concerns functional decomposition and side-effects of actions, of which an agent should be aware if he has actively announced them.

We further propose the following axioms to bridge the gap between mental and communication attitudes. In most cases, however, these will not be needed to reason about communication processes.

14.  $OInt(a_1, A_2, \varphi) \leftrightarrow Bel(a_1, OInt(a_1, A_2, \varphi))$
15.  $OInt(a_1, A_2, \varphi) \leftrightarrow Bel(A_2, OInt(a_1, A_2, \varphi))$
16.  $Op(A_1, A_2, \varphi) \leftrightarrow Bel(A_1, Op(A_1, A_2, \varphi))$
17.  $Op(A_1, A_2, \varphi) \leftrightarrow Bel(A_2, Op(A_1, A_2, \varphi))$
18.  $Op(a_1, a_1, \varphi) \leftrightarrow Bel(a_1, \varphi)$
19.  $OInt(a_1, a_1, \varphi) \leftrightarrow Int(a_1, \varphi)$

According to Axiom 17, (passive) opinions don't necessarily have to be visible, but if we come to believe that someone holds an opinion, this belief is always justified. Axioms 18 and 19 allow us to write (full) belief and intention as special cases of opinion and ostensible intention, respectively.

## 2.4 Communication Attitudes and Communication

Having separated CAs from mental attitudes, the natural next step is to interpret communication in terms of CAs. In the context of this paper, we do not aim at an axiomatisation of a comprehensive set of speech acts like that of FIPA ACL [4], hence giving them an alternative social (but not commitment-based) semantics in terms of CAs. Instead, we provide a minimal set of four communicative acts, and relate CAs directly to actual communications using these communicative acts. *inform* asserts that a particular logical formula currently holds true, *request* declares the intention to make a formula true (note that in our framework, intentionality may include bringing about a certain state or action indirectly "using" other agents), *retract* and *dismiss* cancel a previous statement. Communicative acts are denoted  $sender.act(content, addressee)$ .

The following *trigger axioms* describe the functional relationship between these communicative acts on the one hand and opinions and ostensible intentions on the other (we use a binary *weak until* operator which can be defined as  $\varphi W \psi \equiv_{def} happens((\varphi?)^*; \psi?) \vee \square\varphi$ ):

20.  $(Op) done(a_1, inform(\varphi, A_2))$   
 $\rightarrow (Op(a_1, A_2, \varphi) W done(a_1, retract(\varphi, A_2)))$
21.  $(OInt) done(a_1, request(\varphi, A_2))$   
 $\rightarrow (OInt(a_1, A_2, \varphi) W done(a_1, dismiss(\varphi, A_2)))$

As we have already said, additional rules will be necessary to restrict the use of *retract* and *dismiss* to situations in which it is socially acceptable to retract an assertion or drop an intention. Again, such rules are not to prevent an autonomous agent from uttering certain communication primitives in a given context, but rather define what is considered rational communication and what is not.

## 3 TRUSTABILITY

Kripke-style semantics are widely used to model rational attitudes [see, e.g., 19], and we have pointed out that we consider CAs rational at least to some degree. Yet, they are not very helpful when it comes to determining the actual consequences of someone holding a certain CA (e.g., in terms of reliability). This is crucial because communication is, from an observers point of view, much about decision, behavioural expectation and prediction. As an example for such a *consequentialist semantics* of opinions and ostensible intentions, one might expect that holding a certain ostensible intention means to act "in accordance" with this attitude at least for a limited amount of time, i.e.,

$$OInt(a_1, A_2, done(\alpha_1; \dots; \alpha_n)) \rightarrow \exists h \geq 0. Int(a_1, done(\alpha_1; \dots; \alpha_h))$$

That is, an ostensible intention regarding a sequence of actions implies that the agent *truly* intends a (possibly empty) prefix of this sequence. A similar assumption could be made for opinions. As a matter of rationality in communication, a prefix should truly be intended if it needs to be performed by  $a_1$  alone and under observation by  $A_2$  (i.e.,  $h \geq j$  instead of  $h \geq 0$  if  $\alpha_i = a_1.\alpha_i$  and is observable by  $A_2$  for  $i \leq j$ ). Here,  $h$  can be seen as the *sphere of trustability* of the respective CA, i.e., the length of the largest sequence of actions that are truly intended. Due to the mental opacity of  $a_1$ , the value of  $h$  would in practice have to be determined in a context-dependent way via empirical observation of past ostensible intentions and their consequences in terms of actions actually executed by  $a_1$ . Such a mapping of ostensible to mental attitudes would already provide a consequentialist semantics for ostensible intentions and beliefs, but would come at the price of grounding CAs in mental attitudes. Moreover, it is not at all clear whether and how the extent of the sphere could be computed from past observations, given that intentions denote an earnest self-commitment [17]. To reflect this insight and still allow an observer of a CA to derive information regarding static and dynamic aspects of the world, we propose a consequentialist semantics of CAs that is based on the observer's *trust* that a particular CA does indeed describe (part of) the worlds current state and future evolution.

Trust is commonly defined as a belief that another party will do what it says it will, i.e., be honest and reliable, or reciprocate, i.e., return an advance action for the common good of both, given an opportunity to defect to get higher payoff [see, e.g., 16]. Hence, trust provides an agent  $a_1$  with a means to derive expectations about a certain issue (proposition)  $\varphi$  from CAs regarding  $\varphi$  directed toward him by other agents  $A_2$ . This kind of trust may become crucial when  $a_1$  himself is uncertain about  $\varphi$ , but forced to take impromptu action depending on  $\varphi$ . We give a formal definition.

**Definition 3** *Let  $a_1$  be an agent,  $A_2$  a set of agents,  $\varphi$  a proposition. Then trusting with regard to  $\varphi$  and with degree  $p$  in light of a necessary decision is defined as*

$$\begin{aligned} DTrust_p(a_1, A_2, \varphi) \equiv_{def} & \\ & (happens(a_1.(\alpha \cup \beta)) \\ & \wedge Int(a_1, done(Bel(a_1, \varphi)?; a_1.\alpha \cup Bel(a_1, \neg\varphi)?; a_1.\beta)) \\ & \wedge Bel_q(a_1, \varphi) \wedge (q < 1) \wedge (p + q = 1) \\ & \wedge Op(A_2, a_1, \varphi) \leftrightarrow Int(a_1, done(a_1.\alpha)). \end{aligned}$$

That is, if (i) agent  $a_1$  is forced to choose between actions  $\alpha$  and  $\beta$  and intends to do so depending on whether or not  $\varphi$  holds, (ii)  $a_1$  believes in  $\varphi$  with degree  $q < 1$ , and (iii) some other agents  $A_2$  hold the opinion against  $a_1$  that  $\varphi$ , then  $a_1$  is said to trust  $A_2$  with respect

to  $\varphi$  and with degree  $p$  if he takes action  $\alpha$  if and only if  $p + q = 1$ . A similar definition could be given for ostensible intentions.

Current research on trust in the area of artificial agents mainly focuses on *trust metrics*, which determine how trust and reputation of different dimensions and from various sources are aggregated to form a single notion of trust [see, e.g., 9, 6]. A major weakness of these purely numerical approaches, especially when a precise justification or motivation is required, is the lack of a formal semantics. Existing logical models of trust, on the other hand, including those used in the context of belief revision to assess new information, ground these semantics in mental attitudes of the participating agents, particularly those of the *trustee* [see, e.g., 11, 3]. Like a mentalistic agent communication semantics, this is likely to impede any process by which trust is inferred and used in open environments where agents are mentally opaque. Definition 3 further differs from most existing approaches in that it considers trust at the level of contextualised behavioural expectations or (communicative) assertions rather than entire agents, which results in a more fine-grained and precise model. Since communication is the primary means of interaction among intelligent, autonomous agents, communicative events should constitute the natural, basic subject of trust among such agents. This is in line with the prevailing sociological view of trust as confidence in ones expectations, providing a solution to the problem of acting based on contextualised and gradual (i.e., possibly uncertain) expectations and reducing the inherent complexity of social interaction [12].

While describing precisely (i) what actions need to be taken by an agent who trusts a CA, and (ii) how trust can be recognised ex post, Definition 3 does not provide any additional insights regarding when a CA should or should not be trusted. This underlines the observation of [11] that trust cannot fully be defined as a derived concept. The only available information regarding a decision to trust are past interactions that have taken place in a similar social context or, more precisely, behavioural expectations derived from these interactions. Procedures of arbitrary complexity and sophistication could be imagined to derive trust and hence decisions regarding further action from such expectations. For example, one might want to infer trust (of degree  $p$ ) in agents  $A_2$  regarding proposition  $\varphi$  from gradual belief (of the same degree) that an opinion held by  $A_2$  that  $\varphi$  does indeed mean  $\varphi$ , i.e.,  $Bel_p(a_1, Op(A_2, a_1, \varphi) \rightarrow \varphi) \Rightarrow DTrust_p(a_1, A_2, \varphi)$ . The left hand side of this rule resembles a special case of *adaptive expectations*, i.e., expectations that are revised in case of a disappointment [15]. A particular algorithm by which adaptive expectations can be learned and updated from actual interactions is given in [14]. For reasons of limited space, we refer to the original publications for details. In any case, this process might involve dealing incompatible beliefs, and will hence have to be combined with appropriate mechanisms for belief update or belief revision.

## 4 CONCLUSION

Information in open environments emerges from a potentially large number of competing goals and viewpoints and can be both unreliable and inconsistent. Moreover, a commonly agreed “truth”, or the most basic laws governing reasonable ways to come to an agreement, might not exist. Moving towards a semantically rich formalisation of *opinions* must therefore be a core concern of a strictly communication-oriented paradigm of knowledge representation and distributed AI. This has been the motivation underlying the work described here. Future research will have to concentrate on the application of our approach in real-world domains, and on the integration with the closely related approach of *grounding* in discourse and dialogue games [5].

## ACKNOWLEDGEMENTS

We would like to thank Andreas Herzig for valuable discussions on the topic of this paper and the anonymous reviewers for helpful comments. Part of this material is based upon work supported by the Deutsche Forschungsgemeinschaft under grant BR 609/13-1.

## REFERENCES

- [1] F. Bacchus, *Representing and Reasoning with Probabilistic Knowledge*, The MIT Press, 1990.
- [2] P. R. Cohen and H. J. Levesque, ‘Intention is choice with commitment’, *Artificial Intelligence*, **42**, 213–261, (1990).
- [3] R. Demolombe, ‘Reasoning about trust: a formal logical framework’, in *Proc. of the 2nd Int. Conf. on Trust Management*, (2004).
- [4] FIPA. FIPA communicative act library specification, 2002.
- [5] B. Gaudou, A. Herzig, and D. Longin, ‘Grounding and the expression of belief’, in *Proc. of the 10th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, (2006).
- [6] J. Golbeck and J. Hendler, ‘Accuracy of metrics for inferring trust and reputation’, in *Proc. of the 14th Int. Conf. on Knowledge Engineering and Knowledge Management*, (2004).
- [7] D. Harel, D. Kozen, and J. Tiuryn, *Dynamic Logic*, The MIT Press, 2000.
- [8] A. Herzig and D. Longin, ‘A logic of intention with cooperation principles and with assertive speech acts as communication primitives’, in *Proc. of the 1st Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems (AAMAS)*. ACM Press, (2002).
- [9] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, ‘An integrated trust and reputation model for open multi-agent systems’, *Autonomous Agents and Multi-Agent Systems*, (2006). To appear.
- [10] R. A. Kowalski and M. J. Sergot, ‘A logic-based calculus of events’, *New Generation Computing*, **4**(1), 67–96, (1986).
- [11] C.-J. Liau, ‘Belief, information acquisition, and trust in multi agent systems – a modal logic formulation’, *Artificial Intelligence*, **149**(1), 31–60, (2003).
- [12] N. Luhmann, *Social Systems*, Stanford University Press, Palo Alto, CA, 1995.
- [13] M. Nickles, F. Fischer, and G. Weiss, ‘Communication attitudes: A formal approach to ostensible intentions and individual and group opinions’, in *Proc. of the 3rd Int. Workshop on Logic and Communication in Multi-Agent Systems (LCMAS)*, (2005).
- [14] M. Nickles, M. Rovatsos, and G. Weiss, ‘Empirical-rational semantics of agent communication’, in *Proc. of the 3rd Int. Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, (2004).
- [15] M. Nickles, M. Rovatsos, and G. Weiss, ‘Expectation-Oriented Modeling’, *Engineering Applications of Artificial Intelligence*, **18**(8), (2005).
- [16] S. D. Ramchurn, D. Huynh, and N. R. Jennings, ‘Trust in multi-agent systems’, *The Knowledge Engineering Review*, **19**(1), 1–25, (2004).
- [17] Y. Shoham, ‘Agent-oriented programming’, *Artificial Intelligence*, **60**(1), 51–92, (1993).
- [18] Munindar P. Singh, ‘A social semantics for agent communication languages’, in *Proc. of the IJCAI Workshop on Agent Communication Languages*, (2000).
- [19] Michael Wooldridge, *Reasoning about Rational Agents*, The MIT Press, 2000.