Formalization of Cognitive-Agent Systems, Trust, and Emotions

Jonathan Ben-Naim, Dominique Longin, Emiliano Lorini

Formalization of Cognitive-Agent Systems, Trust, and Emotions

Jonathan Ben-Naim, Dominique Longin, and Emiliano Lorini

Abstract A cognitive agent is an agent characterized by properties that are generally attributed to humans. Cognition is viewed here as a general mechanism of reasoning (in contrast with reactive agents) about knowledge. Such agents can perceive their environment, reason about fact or epistemic states of other agents, have a decision making process, *etc.* This article presents the main concepts used in cognitive agents formalizations, and speak about two particular concepts related to humans: trust and emotion. The language used for cognitive agents is here a logical language because it particularly fits well for both knowledge representation and reasoning formalization. But, even if trust and emotion can be both easily formalized by logical languages, we show that some numerical models are also well adapted.

1 Introduction

To characterize an agent is never easy because a lot of languages can be used, the properties attached to this agent can be various, some concepts may have different names in different contexts, the set of concepts that we need in some context must be different of the set needed in another context, *etc.* In the following, agents are defined as entities having some properties such as: autonomy (they can act without any human action but only with respect to their internal states); reactivity (they can interact with other – human or artificial– agents by using a communication language, or perform some actions that are needed by the environment); pro-activation (they can adopt a behavior following from their goals by taking the initiative); *etc.* As

IRIT-CNRS, Université Paul Sabatier, France,
e-mail: Jonathan.Ben-Naim@irit.fr
e-mail: Dominique.Longin@irit.fr
e-mail: Emiliano.Lorini@irit.fr

it is summarized by Wooldridge [2000], agents are viewed here as computer systems "deciding for themselves what to do in any given situation"

More specifically, in the area of artificial intelligence (AI), the agents properties are often described by using concepts usually associated to humans such as: mental attitudes (belief, knowledge, goal, desire, intention, *etc.*); social attitudes (commitment, common belief or common intention, acceptance, *etc.*), time and action. The properties can also be themselves more specific to humans. We can cite for instance: rationality (in a very wide sense, it means that agents do not act in a contradictory manner: they do not believe both something and its converse, they act with respect to their goals, *etc.*); sincerity (agents do not aim to communicate something they thinks false), *etc.* These properties depend on the context where agents evolve. For instance, is it suitable to have a sincere agent playing poker or an insincere agent supposed to report weather forecasting? Certainly not. So, all the properties used by system designers are selected depending on a particular application.

In the following, we call "cognitive agent system" (or "cognitive system" for short) a system which has a behavior predictable only from its mental attitudes. So, the problem is to determine the mental attitudes that are needed to formalize the properties that we want to attribute to the agents of the system. An advantage of such systems is that they can describe everything, even functional objects (cars, locks, *etc.*). These systems are very popular in AI because they have interesting properties: they are philosophically well-founded, the formal tools are mathematically well defined, the high abstraction level that is used allow to distinguish how something works in the real workd from the general concepts that will used to model it. Finally, these systems have a strong explanatory power (an action mathematically following from both their properties and the agents' mental states that are members of these systems).

In the following, we first speak about cognitive agent systems formalization (Section 2). Such an agent is supposed to be able to: represent its physical environment (including the other agents); represent the manner that it wants this environment evolves; reason about these representations in the aim to perform an action.[1] Logic is a tool that fits very well both this formalizing task and this reasoning task, and this section will only present logical tools (more precisely, modal logics), including three types of operators: belief or knowledge (environment representation), desires, goals, preferences, *etc.* (representation of the wished evolution of this environment), action and time.[2]

Finally, we present two particular concepts strongly related to cognition: trust (Section 3) and emotion (Section 4). We will focus on the cognitive structure of these concepts, that is, on mental states that are necessary to trust or to trigger an emotion. But logic is less appropriate to the representa-

[1] Note that the word *agent* comes from Latin language *agere* and means to act, to do.

[2] These logics are often called BDI logics (for belief, desire, intention). By analogy, we speak also of BDI agents (systems).

tion of their intensity than numerical models. It explains why there are both logical models and numerical models representing trust and emotion. We will give a short overview of these two approaches.

2 Cognitive-Agent Formal Systems

2.1 Short History of BDI Systems

One can say that the story of formal systems as they are today is as long as that of philosophy. Indeed, since Aristotle, philosophy investigated a certain number of concepts: modal logics (logics of necessary and possible), epistemic or doxastic logics (belief and knowledge), deontic logics (obligation, interdiction, permission), temporal, conditional, dynamic logics (explicit or implicit actions), etc.

Our main subject is modal logics, that is, logics including operators that are not truth-functional. So, if $\Box$ is a modal operator, then the formula $\Box\varphi$ (where φ is also a formula of the modal logic) is true independently of the truth-value of φ. This $\Box$ operator can represent beliefs, goals, intentions, etc. For example, if $Bel_i\, sunny$ means that Agent i believes it is sunny, then i can believe it is sunny or not, independently of the weather. (See Chapter **??** of the same volume for more details about modal logics.)

All these formal works, as well as certain others, in particular in philosophy (see [Searle, 1983] and especially [Bratman, 1987]), have contributed to the construction, between end of 80's and beginning of 90's, of the logic BDI of Cohen and Levesque, where: first, intention is defined, in a non-primitive way, from beliefs and goals [Cohen and Levesque, 1990]; and second, the formal framework is also used to characterize the capacities of the agents with regard to communication [Cohen et al, 1990]. One can say that those works have been the corner stone of cognitive-agent systems[3]. Indeed, it suffices to see theories of agents (in particular, those of the language of the agents) as theories of action[4].

Those works have been followed by those of Rao and Georgeff who, based on the logical principals adopted by Cohen and Levesque, have looked forward to a more rigorous formal framework in a temporal logic accompanied with a semantics and an axiomatization [Rao and Georgeff, 1991]. It is worth noting that in those works intention is defined in a non-primitive way. In the same research avenue, we can mention the work of Wooldridge, who introduced

[3] Their paper in *Artificial Intelligence* has received the *AAMAS most influential paper award* in 2008.

[4] This explains by the way the success of the theory of linguistic actions [Austin, 1962; Searle, 1969] in the agent community: in those theories, the language is seen as the accomplishment of actions, facilitating *de facto* the formal union of physical and linguistic actions.

the logic LORA (LOgic of Rational Agent) in [Wooldridge, 2000]. The goal of Wooldridge was not only to formalize an agent architecture of the type BDI, but also its evolution in time.

Concerning french work, we can mention the work of Sadek (see his PhD thesis or KR'92), who, in a formal framework of the same family, defined rationality rules in order to guide the behaviour of a rational agent in a system of rational interactions. By the way, his theory has influenced a language of agent communication (agent communication language or ACL) that became an international reference, which has been used or gave rise to numerous works in the agent community: the FIPA language[5].

In the mid 90's, more operational languages appeared, in the sense that the goal is not only to have a logical formalism able to capture the concepts useful to construct the agent systems of interest, but also to implement them. So, BDI systems formalized in situation calculus appeared (see for example the works of Shapiro, Lespérance, and Levesque in Toronto). Programming languages based on primitives of the BDI type also appear: one can mention e.g. GOLOG or ConGolog. This community gave rise to what can be called nowadays cognitive robotics, whose laboratory of the same name in Toronto is the most prominent representative.

Certain formalisms also aims at describing normative systems, that is, systems where the agents have not only to consider what they believe (or know) and what are their goals, but also what they must do. This aspect uses (also also inherits theoretical questions from) deontic logic. For example, we can mention the BOID architecture (where O represent the obligation component of the BDI system) of van der Torre *et al.* (see e.g. the paper published in AGENTS'01).

Next, BDI systems not only manipulate mental attitudes (in addition to time and/or action), but also social concepts or external constraints. Obligation can be seen as an internal norm (it is then formalized by an operator indexed by an agent or a group of agents), or as an external law every agent must obey (it is then formalized by a non-indexed operator).

By the end of 90's, the BDI systems, as they are then formalized, are heavily criticized, because they are based on strong hypotheses about mental states, in particular sincerity. So, in FIPA for example, an agent believes everything it is told by another agent, because it always assumes the latter tells the truth.

To avoid this problem, certain works describe the effect of a linguistic action by separating what the speaker wants to mean from what the listener believes on the basis of hypotheses made by the latter about the sincerity and competence of the former. Other works looked forward to alternative concepts allowing us to free us from those hypotheses about the internal states of the agents. For example, there are numerous works on social commitment aiming at capturing the public commitment of an agent generated by what that agent

[5] http://www.fipa.org/repository/aclspecs.html

says. For instance, when someone says something, he (or she) is committed to the truth-value of that proposition: he could not say he did not said it, and cannot say or do something that opposes what he said (see e.g. the work of Singh [Singh, 1999] and de Colombetti in Switzerland). Nevertheless, those approaches also have drawbacks: other hypotheses are made (for example, the public aspect of linguistic actions and the fact that they are correctly interpreted by their targets). In addition, it is not obvious that this concept is devoid of links with the mental states of the committed agents.

Finally, this notion has not been studied in a satisfactory way as a non-primitive concept[6], despite the fact it apparently contains a normative and conventional component, as well as violation condition. In such circumstances, those approaches are almost not BDI systems, since they do not involve mental states: there is an intuitive link, but it has to be formally established.

Other traditional concept have been confronted to that problem, e.g., common belief. The latter is generally defined as the infinite conjunction of the alternative beliefs between agents. For example, if there is common belief between agents i and j about φ, then i believes φ, j believes φ, i believes j believes φ, j believes i believes φ, i believes j believes i believes φ, etc.

Thus, the problem in an implemented system is to decide whether there is common belief without having access to the minds of the agents. At best, we can construct a subjective notion of common belief, i.e., the fact that an agent believes there is common belief (maybe it is not the case). A number of philosophical works are related to this question (see e.g. [Gilbert, 1989]). They led to notions like acceptance (see e.g. [Lorini et al, 2009]).

In parallel, certain prior AI problems have been transfered to the BDI framework and gave rise to a rich literature: the frame problem (how to describe environment in a concise and exhaustive fashion?), the problem of characterizing actions (what are the necessary and sufficient conditions to execute a given action?), the problems of revision (how to have an agent's knowledge evolves with time?) and action ramification (how to describe the impact of an action on the domain, including the mental states of the agents). For example, the advent of BDI systems was followed by the problem of revising mental states (see e.g. [van der Hoek et al, 2007]).

More recently, this problem has become the heart of a branch of the domain: dynamic epistemic logics (see below). Put simply, the goal is to integrate into the semantics of these logics the fact that the beliefs (or knowledge) of an agent can evolve: that agent can change his mind, learn that certain propositions are true, learn that others are false, etc. At the cost of certain technical constraints, the logics of public announcements give an adequate answer to the hard question of mental-states evolution. For an overview on that subject see e.g. [van Ditmarsch et al, 2007].

[6] that is, a concept constructed from lower-level concepts.

Finally, agent testbeds have been developed, like e.g. AgentSpeak by Rao, Jason by Hübner and Bordini, or 2APL by Dastani. Those testbeds allow the implementation of agents and multi-agent systems, but do not yet exhaust all the expressive power of the BDI logics. In particular, they are not equipped with a complete set of boolean operators and do not use theorem provers, which by the way already exist for certain (families of) well-known logics.

Concepts proposed in the domain of BDI systems have also been used in other domains of AI. For example, this is the case of argumentation, where e.g. Amgoud used argumentation methods to generate desires and plans in an autonomous agent [Amgoud and Rahwan, 2006] (see also Chapter **??** of the same volume).

2.2 Basic Concepts

In what follows, we present the basic concepts generally used in the formalization of cognitive-agent systems in terms of mental states. Of course, all systems do not use all concepts simultaneously, because the way an agent system is characterized depends on the domain of that system.

As soon as we need nested operators, modal logics are particularly adequate, because a formula of a modal logic in the range of a modal operator forms a new formula of that logic. So, we can have an arbitrary large degree of nestedness in the formulas of the object language. This property is particularly important in the domain of cognition, because we can have beliefs about almost anything, including other beliefs: Agent i believes Agent j believes Agent k believes Agent i believes p, etc. (see Chapter **??** on knowledge representation of the same volume).

2.2.1 Belief Operators

The notion of belief has been deeply studied in the domain of doxastic and epistemic logics, since the early 60's (see [Gochet and Gribomont, 2006] for an exhaustive overview). This is probably one of the most studied notion in Logic, in all its forms (classical logic, modal logic, with or without degrees representing the strength of the beliefs or knowledge of an agent[7]).

A commonly used logic is the propositional modal logic without degrees where "Agent i believes φ is true" is denoted by $Bel_i\,\varphi$, where Bel_i (for every agent i) is called the modal operator of Agent i's beliefs, and where φ is some formula. Traditionally, the fact that $Bel_i\,\varphi$ is true in a certain world w_0 is interpreted as the fact that φ is true in all worlds that are accessible from

[7] In the present work, we only consider qualitative approaches to the notion of belief. We do not discuss the quantitative approaches formalizing degrees of belief (see e.g. [Laverny and Lang, 2005]).

w_0 according to Agent i. Note that i has no certainty that the real world belongs to this set of epistemic worlds (i may be wrong). To represent this, the semantics includes an accessibility relation for every agent. So, the fact that i believes φ is true in the real world w_0 is denoted by $w_0 \Vdash Bel_i\,\varphi$. Semantically, this means φ is true in all worlds that are accessible from w_0 via the relation corresponding to i and denoted by $\mathscr{B}_i$.

There is a consensus in the literature that the logic of beliefs in the normal modal system KD45 [Chellas, 1980], even though this logic constitutes an idealization of certain principles. For example, this logic assumes an agent instantly knows all beliefs implied by its own (omniscience) and it is conscious of all those beliefs (positive introspection). Nevertheless, those criticisms are mitigated by the fact that they constitute idealizations (not aberrations), which are not necessarily counter-intuitive for an artificial agent.

Fig. 1 shows the semantics of the belief operator of agent i. The set of all worlds that are accessible from w_0 is denoted by $\mathscr{B}_i(w_0)$, where $\mathscr{B}_i$ is the accessibility relation between worlds for Agent i and is graphically represented by arrows.

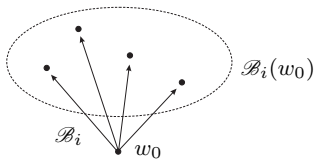

Fig. 1 Kripke semantics of the operator Bel_i

2.2.2 Temporal Operators

There are many temporal logics, depending on the way one wants to represent time (ramified or linear, with or without explicit temporal indexes, etc.). Temporal logics are relatively well-studied in the domain of modal logics and theoretical computer sciences [van Benthem, 1991]. There semantics is based on transition relations between possible states and are thus equivalent to (potentially infinite) automates (see Chapter **??** of the same volume for more detail about temporal reasoning).

Here, we focus on a very simple notion: linear time. Since this notion is combined with the beliefs of the agents, this means that the latter are not about epistemic worlds, but about linearly-ordered sets of worlds called "stories". This allows us to simulate a tree-based nature of time, since each story corresponds to a development of future events (the agent believes possible).

For example, Fig. 2 represents the four stories believed by Agent i. The dots on the stories represent the present moment and the dashes the past and

future moments. So, the agent right now believes that p is true ($Bel_i\, p$); it consider the possibility that r is right now true but becomes false thereafter ($\neg Bel_i \neg(r \wedge F\neg r)$); etc.

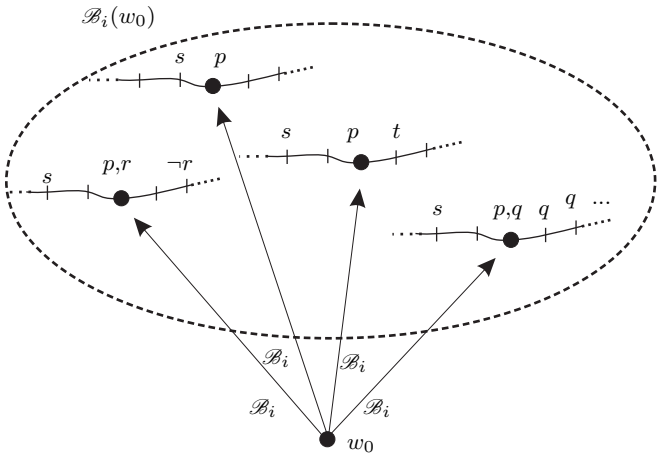

Fig. 2 Linear time and epistemic worlds

We can define the operators H and P (about the past) in the same way we defined the operators G and F.

Technically, time is defined in a modal logic of linear time of Type S4.3$_t$ (see [Burgess, 2002] for more details). Nevertheless, those operators can be semantically defined with a tree-based structure (which is by the way what is done in [Rao and Georgeff, 1991]).

Finally, we sometimes use the two operators X and X^{-1} such that $X\varphi$ means "φ will be true right after the present moment is the considered story" and $X^{-1}\varphi$ means "φ was true right before the present moment in the considered story". Obviously, there exists formal links between those operators and the temporal ones defined previously.

2.2.3 Goal Operators

The notion of goal has been widely studied in the literature and has been used in very different senses (see e.g. the notion of goal in Cohen and Levesque [Cohen and Levesque, 1990] or Rao & Georgeff [Rao and Georgeff, 1991], the notion of choice in the PhD thesis of Sadek or KR'92). We focus on the notion of *chosen goal* (or *preferred goal*), with regard to the coherent subset of proposition the agent wants to make true. The primitive operators of goal are denoted by $Choice_i$ (where i ranges over all agents) and $Choice_i\, \varphi$ means that "Agent i right now choose to make the goal φ right now true".

There is no restriction on the formula φ, so it can represent the present state of affairs. This is the difference with the operators of goals to be achieved (abandoned when the desired state of affairs comes true) or to be maintained (an agent looks forwards to keep a certain state of affairs true). As we did it with beliefs, we interpret $Choice_i\,\varphi$ in a world w_0 as the fact that φ is true in all the preferred world of the agent from w_0. Most generally, goals are partial pre-order, but, for the sake of simplicity, we do not consider this point: we focus on coherent non-ordered binary goals.

A difficult and non-studied question is the following: how those goals emerge? From a cognitive point of view, it looks like they emerge from a deliberative process about more primitive attitudes: desires, ideals, and imperatives (see [Rao and Georgeff, 1991; Conte and Castelfranchi, 1995; Castelfranchi and Paglieri, 2007]). The set of goals we characterize is the one obtained from a process of selection of ideals and desires. It is meant to resolve conflicts between those two concept and to eliminate impossible cases. Then, the chosen goals of an agent satisfy the two following fundamental rationality principles: they are consistent (an agent cannot choose two contradictory goals); the chosen goals are related to the beliefs of the agent that chose them. In [Cohen and Levesque, 1990], the relation between beliefs and goals is an inclusion relation: if an agent right now believes φ is true, then it necessarily right now has φ as a goal (this notion is called *strong realism*). We can also impose a relation of *weak realism*, where it is only required that there is a non-empty intersection between the epistemic worlds that are possible and those that are preferred.

Some recent works aim to explain the goals building process by the way of desires. Desires and goals are often combined (see for instance [Dubois et al, 2017]).

2.2.4 Ideals

There exist many normative systems in logic with very different characteristics, more or less complex, adapted to a class of problems or another. Those norms may have different origins: state laws, institution rules, moral (be it religious or not), etc.

Certain particular norms, specific to a given agent, are called ideals. We introduce a new set of operators such that $Idl_i\,\varphi$ means: "φ is an ideal state for Agent i". This means that i gives an order to itself, a kind of "must make true" for φ (when φ is false at the present moment) or "must keep true" (when φ is already true) [Castaneda, 1975].

There are different ways to explain how a state φ becomes an ideal state for a certain agent. A possible explanation is that ideals are just social norms that have been internalized (or adopted) by this agent [Conte and Castelfranchi, 1995]. Assume an agent believes in a certain group (or institution) there is a certain norm (e.g. an obligation) saying that a state φ must be true, whilst

the agent sees itself as a member of that group. In such a case, the agent adopts this external norm (that does not originate from the agent and has not yet been acknowledge as a norm by the agent) and that norm becomes an ideal for that agent. For example, if Agent i believes in France, it is obligatory to pay taxes and that agent considers himself (or herself) as a French citizen, then he adopts this obligation and pays his taxes.

Semantically, the ideals are represented from the possible worlds considered as ideal by the agent having internalized those ideals. There is no particular relation with the other operators, besides belief, if we assume an agent is conscious of its ideals (see Chapter **??** of the same volume for more details about normative operators). (See also [Gabbay et al, 2013] pour for more details about normative and deontic systems and [Berreby et al, 2015; Lorini, 2016] about moral systems.)

2.2.5 Explicit Action

When one tries to define "Agent i is capable of executing Action φ", one has to consider logics of action (see Chapter **??** of the same volume on reasoning about action and change). Generally speaking, those logics formalize actions with state-transition systems. There are essentially to schools of thought, one where action is explicit and one where it is implicit (see the next section).

The main logic of explicit action is propositional dynamic logic (PDL), which studies the interact between an action and its effects [Harel et al, 2000]. It has been shown (e.g. in [van Linder et al, 1998]) that dynamic logic is particularly adapted to the characterization of the concepts of capacity and power. There is a rich literature on the integration of dynamic logic into logics of beliefs and goals (see e.g. epistemic dynamic logic [Baltag and Moss, 2004] or doxastic dynamic logic [Segerberg, 1992, 1995]).

PDL distinguishes between actions like α and formulas like φ and ψ, and its set of non-logical constants is constructed from those two categories. The formula $\mathit{After}_{\alpha}\,\varphi$ expresses the fact that φ will be true after any possible execution of Action α. So, $\mathit{After}_{\alpha}\,\bot$ means α is not executable [8].

Several extensions have been proposed where an agent is added to the arguments of the PDL operators. In such extensions, the formula $\mathit{After}_{i:\alpha}\,\varphi$ means that φ is true after any possible execution of Action α by Agent i. For any action α and agent i, $\mathit{After}_{i:\alpha}$ is an action modal operator.

Semantically, action is treated as a transition from a real world to a set of other real worlds (certain semantic constraints can force this set of worlds to be a singleton). Fig. 3 represents this transition.

In DEON'2008, Lorini and Demolombe have augmented the PDL language with the operators $\mathit{Does}_{i:\alpha}$, where $\mathit{Does}_{i:\alpha}\,\varphi$ means "Agent i is about to

[8] Besides BDI logics, the operator After_{α} is often denoted by $[\alpha]$.

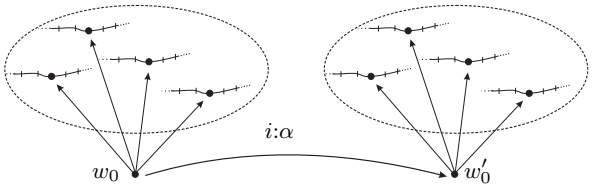

Fig. 3 Transition from the world w_0 to the world w'_0 *via* the execution of the action $i{:}\alpha$

execute Action α and thereafter φ will be true". This allows us to speak about what an agent can do ($\neg After_{i:\alpha}\,\bot$) and what an agent does ($Does_{i:\alpha}\,\top$).

2.2.6 Implicit Action

Action is implicit in logics of agency, which study the interaction between an agent and the effects caused by it. The peculiarity of those logics is that they do not represent the actions that caused the effects (only results matter).

For example, in the logic *STIT* [Belnap et al, 2001], actions are formalized by formulas involving an agent and speaking about the effects caused by that agent. So, the action described in "i buys the product p" is formalized by the following formula of agency: "i sees to it that Product p is bought by Agent i".

Formulas of agency are of the form $STIT_i\,\varphi$, which means "The action chosen by Agent i at the present moment ensures that φ is true, independently of what the other agents do". In short, "i sees to it that φ". The modal operator $STIT_i$ is called the operator of agency.

2.2.7 Dynamic of Mental States

Last years, a certain number of researchers working in the domain of logics for autonomous-agent formalization and in multi-agent systems have proposed logics for the dynamic of mental states. They belong to the large family of dynamic epistemic logics (DEL), see e.g. [Ditmarsch et al, 2007]. DEL is a term used in a very large sense to include dynamic extensions of logics of belief and knowledge, but also logics of preferences and norms (deontic logics) [Baltag and Moss, 2004; Kooi, 2007; van Benthem and Liu, 2007]. In those logics, modal operators are introduced to describe the effects, on the mental states of the agents, of various types of informative events (transmission of public or private messages, orders, etc.).

Here, we consider the most known dynamic epistemic logic, namely public announcement logic (PAL) [Ditmarsch et al, 2007]. Informally, a fact p is publicly announced if and only if: every agent learns that p is true; every

agent learns that every agent learns that p is true; every agent learns that every agent learns that every agent learns that p is true, etc., up to infinity. In the PAL logic, public announcements are events that update the beliefs and knowledge of the agent: the role of a public announcement is, first, to reduce the set of possible worlds to the worlds where the publicly announced fact holds, and second, to restrict the epistemic accessibility relations to those worlds. PAL uses the notation $p!$ for the public announcement of p, and introduce modal operators of the form $[p!]$ to describe the effect of a public announcement on the mental states of the agents: the formula $[p!]q$ means that q will be true after the public announcement of p. We take below an example to illustrate those dynamic operators.

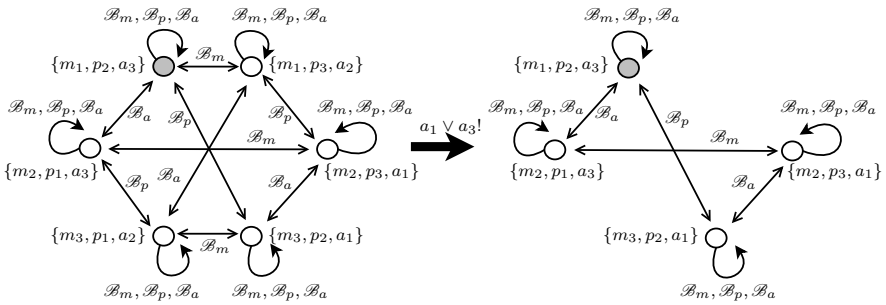

Fig. 4 Example of cards

Marie, Paul, and Alice are seated around a table on which are laid three cards. The cards are face down, but, on every card, is written a distinct number between 1 and 3. So, the cards can be called Card 1, Card 2, and Card 3. Marie, Paul, and Alice take one card, each. We assume Marie took the card 1 (denoted by m_1), Paul the card 2 (denoted by p_2), and Alice the card 3 (a_3). Each player confidentially looks at his (or her) card and put it of the table face down. Therefore, each player only knows the number written on his card.

In Fig. 4, the model on the left represents the beliefs of Marie, Paul, and Alice in the initial situation. There are 6 possible worlds and the one in grey is the real one. The arrows represent the accessibility relations $\mathscr{B}$ between epistemic worlds, for each player. For example, in the real world, Marie considers as possible the world where Marie has Card 1, Paul Card 2, and Alice Card 3, as well as the world where Marie has Card 1, Paul Card 3, and Alice Card 2. So, in the real world, Marie has no certainty about the card distribution.

Assume it is publicly announced that Alice has a card with an odd number. This announcement is represented by the event $a_1 \vee a_3!$ (Alice has Card 1 or Card 3). In Fig. 4, the model to the right of the arrow represent the beliefs of

Marie, Paul, and Alice after this announcement. Thanks to the latter, Marie learns that Paul has Card 2 and Alice Card 3. Indeed, the effect of the public announcement is to reduce the set of possible worlds to those where Alice has an odd card and to restrict the accessibility relations to those worlds. So, in the real world, after the public announcement, Marie knows the card distribution: Marie has Card 1, Paul Card 2, and Alice Card 3. This fact is represented by the formula $m_1 \wedge p_2 \wedge a_3 \wedge Bel_m(m_1 \wedge p_2 \wedge a_3)$, which is true in the real world of the model on the right. In contrast, the public announcement does not make Paul and Alice learn anything: after the public announcement they still have no certainty about the card distribution.

Up to know, we gave an overview of the concepts related to cognitive-agent systems and a way to formalize them. In what follows, we present two particular complex concepts that can be described in terms of mental states, time, and action. Cognitive-agent systems are thus very adapted to the formalization of these two concepts. Nevertheless, the latter are also formalized in more numerical ways and, in what follows, we give an overview of this.

3 Formalization of Trust

Trust systems (or trust models) are used in certain multi-agent systems to help users to choose the agents to interact with. Indeed, agents may be incompetent or malicious. But, the agents are typically so numerous that it is impossible for a central authority to test them all. Consequently, the goal of a trust system is to evaluate the agents on the basis of relations between them. More precisely, for a user u, the evaluation of the peers of u is based on two kinds of information:

- the result of past interactions between u and the other agents;
- the feedbacks other agents have provided about their peers.

The value (a score, a position in a ranking, etc.) of an agent a can naturally be seen as the trust of u in a.

Trust systems can be motivated by several large-scale applications where no central authority can test all agents. As examples, we can mention: e-commerce (Ebay, Amazon, etc.), large wikis (Wikipedia, Planetmath, etc.), social networks (Facebook, Tweeter, etc.), webpages and hypertext links, papers and citations.

Various trust systems have been developed. To validate and compare them two kinds of approaches are possible: a theoretical and an experimental one. The first approach consists in establishing desirable properties (or axiom,

postulates) that a trust system could satisfy. The second approach consists in developing a testbed where different trust systems can compete.

As far as we know, there are two kinds of trust systems: logic-based systems (essentially modal-logic-based systems) and numeric systems. Two position papers that cover a large number of models are for example [Sabater and Sierra, 2005] and, more recently, [Pinyol and Sabater-Mir, 2013].

3.1 Logic-based Trust Models

In the logical approach, the goal is to characterize the notion of trust in a certain formal language. Similarly, the objective is to formalize in such a language the notion of trusting someone, as well as the mental state of an agent trusting someone.

One of the main models of trust is the cognitive one from Castelfranchi et Falcone (denoted by C&F) [Castelfranchi and Tan, 2001]. Contrary approaches that are more computational, the C&F model is more than subjective probabilities updated in the light of direct interactions with the *trustee* (the agent to be trusted) and feedbacks from interactions between the trustee and other agents.

Informally, the C&F model defines trust as an personal belief of the *truster* (the agent that has to decide whether or not to trust the trustee) that the trustee is reliable with regards to various aspects (capacity, intention, readiness, etc).

According to C&F and the analysis conducted in [Herzig et al, 2010], the notion of trust is based on four components: a truster i, a trustee j, an action α of j, and a goal φ of i. According to their definition, "i trusts j that j will perform α in order to achieve φ" if and only if: φ is a goal of i; i believes j is capable of performing φ; i believes that performing φ will makes φ true; and i believes that j intends to do α.

For example, assume i trusts j to send a certain product p in order to possess p. Then: possessing p is a goal of i; i believes j is capable of sending p; i believes sending p will make him (or her) possessing p; and i believes j intends to send p.

In other words, trust is formally defined as follows:

$$\mathit{Trust}(i,j,\alpha,\varphi) \stackrel{def}{=} \mathit{Goal}_i\,\varphi \wedge \mathit{Bel}_i\,(\mathit{Capable}_j(\alpha) \wedge \mathit{After}_{j:\alpha}\,\varphi \wedge \mathit{Intend}_j\,(\alpha))$$

where every operator used above is either a basic one or a compound one defined from the basic ones (*cf.* Section 2.2):

- $\mathit{Goal}_i\,\varphi \stackrel{def}{=} \mathit{Choice}_i\,F\varphi$ means "Agent i chooses to make $F\varphi$ true at the present time";

- $Capable_j(\alpha) \stackrel{def}{=} \neg After_{j:\alpha} \bot$ means "Agent j is capable of executing Action α if and only if α is already executable";[9]
- $Intend_j(\alpha) \stackrel{def}{=} Choice_j\, Does_{j:\alpha} \top$ means "Agent i intends to execute Action α if and only if executing α (right here, right now) is a chosen goal of i".

A relatively recent paper allowing an agent to reason about its trust model, by providing a method for incorporating a computational trust model into the cognitive architecture of the agent is [Koster et al, 2013].

We turn to approaches where the notion of trust is not based on modal logic, but more numeric objects.

3.2 Numerical Models of Trust

Previously, trust was seen essentially as a particular belief of the truster about certain aspects of the trustee. Depending on whether i trusts j or not about a proposition φ, i was in position to decide whether or not to believe what j says about φ.

The situation is similar with numeric approaches. The first question is to decide how to represent trust in a numeric way. Various solutions have been proposed, for example, trust can be represented by a number, an interval, or a fuzzy interval.

First, trust can be represented by a simple number. One of the first approaches of this kind is [Marsh, 1994]. Another important approach is that of *Pagerank* [Page et al, 1998], the system at the basis of the well-known Google search engine. More precisely, a webpage can be seen as an agent and a hypertext link from x to y as a positive feedback. Pagerank associates every agent with a real number between 0 and 1 on the basis of these feedbacks. Theses numbers can be seen as the degrees of trustworthiness of the agents.

It is worth noting that Pagerank evaluates the trustworthiness of an agent for an external user. Most trust systems evaluates the trustworthiness of an agent for another agent x. In such a case feedbacks from direct interactions with x are obviously more important than feedbacks from interactions where x is not involved.

A relatively exhaustive study of questions related to Pagerank and its alternatives can be found in e.g. [Langville and Meyer, 2005]. A version of Pagerank adapted to peer-to-peer systems as been constructed in [Kamvar et al, 2003].

In certain approaches, an agent is either trustworthy or not, and a model can associate an agent x with a number indicating the probability that x is

[9] One could think that this should be a sufficient but not necessary condition. Indeed, it suffices that Agent i believes Agent j will be capable of executing Action α in time to achieve Goal φ. Nevertheless, it is worth noting that we formalize a notion of trust "right here, right now", not a notion of potential trust.

trustworthy. In other approaches, a model can associate an agent x with a number indicating the degree of trustworthiness of x. In other words, depending on the model, the same number x is associated with can mean different things. For example, assume x is associated with 0.5. It can mean that x perfectly achieves one goal out of two, as well as x achieves every goal half-successfully.

Concerning links between trust and other important notions, it is described in e.g. [Osman et al, 2015] how trust models can be used to distinguish between good and bad advices. Finally, a paper describing how the notion of trust can be integrated with those of negotiation and argumentation is e.g. [Bonatti et al, 2014].

3.3 Applications of Trust Systems

We present six examples of multi-agent systems where a user (be it an external entity or an internal agent) needs an evaluation of the trustworthiness of the agents:

E-commerce (Ebay, Amazon, ...).

The agents are the buyers and sellers. A user has to choose the agents to make transactions with. But they are numerous, generally unknown to him (or her), far from him, and some agents are malicious or incompetent. So, the user needs an evaluation of the agents. After each transaction, the buyer can rate the seller, and vice versa. So, a trust system can exploit these ratings to construct an evaluation. We can globally admit that the more an agent is trustworthy, the more he tends to provide honest and accurate feedbacks. The same goes for the buyers. So, in case of cycles the trustworthiness of an agent x depends on that of an agent y, and vice versa, which makes the evaluation hard to construct.

Large wikis (Wikipedia, Planetmath, ...).

The agents are the contributors of the wiki, that is, those that create, delete, or modify articles. A user has to choose to trust or not the contributions and thus needs an evaluation of the contributors. It is easy to imagine how to modify a wiki so the contributors can provide opinions about their peers, in particular when they participate in long debates about controversial issues. A trust system could exploit these opinions to construct an evaluation. We can admit that the more an agent provides serious contributions, the more he

(or she) tends to provide serious opinions about its peers. So, again opinion cycles constitute a difficulty.

Social networks (Facebook, Myspace, ...).

The agents are the persons, applications, etc. registered in the network. A user has to choose the agents to establish a formal link with. Such a link gives access to all sorts of personal information about the user. But, some persons or applications are malicious. A friendship link between a and b can be seen as the fact that a recommends b as an honest agent, and vice versa. Those links can exploited to evaluate trustworthiness. There are recommendation cycles and the more an agent is honest, the more it provides honest recommendations.

Web pages and hypertext links.

The agents are the web pages. A hypertext link from a page a to a page b can be seen as a recommendation, that is, as the fact that a provides an opinion that the content of b is important. There are cycles and the more a page a has an important content, the more the hypertext links contained in a are important.

Papers and citations.

An agent is a paper or an author. A citation relation from a paper x to a paper y can be seen as the fact that x supports y. Similarly, an authorship relation between a paper x and an author a can be seen as a support relation for a. The are no relation cycles, but the more a paper x is trustworthy, the more the citation or authorship relations coming from x are important.

Entity-key bindings and certificates.

In the systems based on public key certificates, there are entities willing to send messages to other entities. Since an entity can listen messages that are not intended for it, they are encrypted and decrypted with keys. So, the system generates a set of keys such that there exists a function f transforming any key K into a key $f(K)$ such that the following holds:

(a)$f(K)$ is the unique key that can decrypt the messages encrypted with K, and it can decrypt only these messages;
(b)the converse is true, that is, K is the unique key that can decrypt the messages encrypted with $f(K)$, and it can decrypt only these messages.

Next, a set of bindings is published. A binding is a pair $\langle E, K\rangle$ where E is an entity and K a key. Such a binding represents a claim that E is the unique entity that knows $f(K)$. If it is indeed the case, then we say that $\langle E, K\rangle$ is valid. So, to send a confidential message to an entity, it suffices to find a binding containing it, and use the corresponding key. By (a), only this entity will be able to decrypt the message. The problem is that a malicious entity F can publish a false binding $\langle E, K\rangle$. In other words, E does not know $f(K)$, but F does. So, if this false binding is used, then F can listen some messages intended for E and decrypt them.

To counter this, a set of public key certificates is published. A certificate is a pair $\langle D, S\rangle$, where D is a quadruplet of the form $\langle E, K, E', K'\rangle$ and S is a digital signature, that is, S is supposed to be the result of encrypting D with $f(K)$. Such a certificate represents a claim that E supports the validity of $\langle E', K'\rangle$. Again, the problem is that false certificates can be published. However, it is possible to formally check that the certificate $\langle D, S\rangle$ was created by an entity knowing $f(K)$. By (b), it suffices to decrypt S with K and then check that the result is indeed equal to D. Only the certificates that pass this test are considered.

Now, we can explain the link with trust systems. An agent is a binding $\langle E, K\rangle$. A user is an entity E that has to choose valid bindings before sending messages. A certificate $\langle\langle E, K, E', K'\rangle, S\rangle$ can be seen as the fact that $\langle E, K\rangle$ supports the validity of $\langle E', K'\rangle$. These support links can be exploited to evaluate the validity of the bindings. Finally, the problem of evaluating the validity (or trustworthiness) of the bindings is difficult in particular because there are cycles of support links and the following holds: if a binding $\langle E, K\rangle$ is valid, then E is the unique entity knowing $f(K)$, thus the certificate $\langle\langle E, K, E', K'\rangle, S\rangle$ was created by E, i.e., this certificate is authentic, so we should attach more importance to it.

4 Formalization of Emotions

There is a rich literature on emotions, be it in philosophy [10] [Gordon, 1987], psychology [Lazarus, 1991; Ortony et al, 1988], economy [Loewenstein, 2000], or cognitive sciences [Lane and Nadel, 2000].

In computer sciences, emotions play an important role in multi-agent systems at different levels. Much work focus on the modelization of facial and gestural results of emotions with animated conversational agents (ACA) (see e.g., [Gratch and Marsella, 2005; Pelachaud, 2009]). ACA also use models of emotions to represent those of the users, to show their affective states, or a particular personality.

[10] Plato clearly establishes a distinction between reason, passion, and desire.

The goal is to make such agents so realistic that users have the impression to interact with other humans. First, this goal assumes a great realism in the expressive aspects of the agents (facial and corporal movements, intonations, verbal expressions, etc.). Second, it is necessary, for the agents, to be able to recognize and take into consideration user's emotions in their of reasoning (as well as their artificial own). So, agents can speak and act in a most adequate fashion.

Emotions are fundamental to have natural and optimal interactions between agents and users, because nowadays it is known that we constantly communicate information about our emotional states (be them real or not) without explicating them. For example, a "Hello!" accompanied with a smile constitutes a common and short way to express your greetings to someone and to tell him (or her) you are happy to see him (which could be explicated by "Hello, I am happy to see you").

4.1 Logical Formalization of Emotions

Concerning formal models of emotions, we look forward to construct logical frameworks in order to formalize certain specific emotions, their properties, the links between them, etc. (see e.g., [Adam et al, 2009; Turrini et al, 2010]). The main objective is to take advantage of logical methods to rigorously specify how to implement emotions into an artificial agent. The design of systems containing such agents (capable of reasoning and expressing certain emotions) can benefit from the fact that logic is a tool particularly adapted to the notion reasoning and forcing the designer to disambiguate the different dimensions of emotions (identified in different psychological models of emotions).

Generally, logical definitions of emotions characterize cognitive structures of emotions, rather than emotions themselves. According the theories of cognitive evaluation [Lazarus, 1991], the cognitive structure of an emotion is the configuration of the mental state of an agent when it (artificially or not) feels that emotion. The cognitive structure is just a part of the affective phenomenon. In the sequel, we use the word "emotion" for "the cognitive structure of an emotion".

We distinguish between simple emotions and what we call complex emotions [Adam et al, 2011; Lorini and Schwarzentruber, 2011]. The former are those that can be described only with mental attitudes like beliefs, goals, or ideals. The latter are those requiring more complex reasonings like counterfactual conditionals: "I could have made φ true, whilst it is actually false". In that sense, complex emotions are associated with counterfactual reasonings about norms, responsibilities.

For example, the fact that agent i feels joy about Fact φ may be expressed as follows:

$$Joy_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Choice_i\,\varphi$$

According to this definition, Agent i feels joy about φ if and only if i believes φ is true and wish φ to be true. For example, Tom feels joy about a certain test, because he thinks he successfully passed it and it is what he wishes. So, Tom is happy because he believes the state of affairs is as he wishes. Joy has a positive valence, that is, when it is felt, it is associated to a state of affairs corresponding to desires. This is not the case of sadness for example whose state of affairs does not correspond to desires.

Concerning complex emotions, we restrict ourselves to those related to the notion of responsibility (be it that of the agent feeling the emotion or another one). The responsibility of Agent i for the fact that φ is true can be defined as follows: φ is true and i could have made φ false. More formally:

$$\mathbf{Resp}_i\varphi \stackrel{def}{=} \varphi \wedge \mathbf{Cd}_i\neg\varphi$$

Here $\mathbf{Cd}_i$ (i could have made) is a basic operator of the formal language, but can be defined from the implicit action operator STIT (for more details, see [Lorini and Schwarzentruber, 2011].)

So, when Agent i is responsible for the fact that φ is true, whilst i has $\neg\varphi$ as goal, i feels regret (see e.g., [Zeelenberg et al, 1998]). More formally:

$$\mathit{Regret}_i\varphi \stackrel{def}{=} \mathit{Goal}_i\,\neg\varphi \wedge \mathit{Bel}_i\,\mathbf{Resp}_i\varphi.$$

Other emotions can be defined in the same way. Emotions constitute a growing domain, because computer science does not have yet exhausted all their possibilities. Existing and implemented systems can often be reduced to simple labels that can be activated or deactivated. Formal models based on logic force designers to explicate the nature of emotions and thus to better understand them.

4.2 Numerical Models of Emotions

There exist numerical models of emotions that study the quantitative aspects of those affective phenomena. For example, El-Nasr et al. [El-Nasr et al, 2000] proposed a numerical model of emotions called FLAME (Fuzzy Logic Adaptive Model of Emotions) based on fuzzy logic. The main contribution of this work is a quantification of the intensity of emotions, from appraisal variables like desirability or probability. For example, based on the psychological model of emotions of Ortony, Clore and Collins [Ortony et al, 1988], in the model FLAME, the intensity of hope with regard to a certain event depends on the degree of desirability of that event and its subjective probability. More recently, several researchers in AI have augmented formal models of emotions with quantitative aspects. For example, Meyer et al. [Steunebrink et al, 2008] proposed a model describing how the intensity of emotions decreases

with time. Lorini [Lorini, 2011] proposed a systematic study of the intensity of emotions on the basis of expectations (hope, fear, disappointment, relief) and the relation between those emotions and the mechanism of belief revision of a cognitive agent.

There also exist numerical models of emotion where the latter is represented by a vector whose numbers correspond to components of emotion. For example, Mehrabian captures mood by a vector representing pleasure, excitation, and dominance (i.e., the capacity of an individual to dominate a stimuli). In other words, mood depends on the values of those three components. We can also mention works on the robot with human-like head WE-4R constructed in the university of Waseda (Japan) by Hiroyasu Miwa and his team. The model of emotion is a space-oriented vector calculated from three components: pleasure, activation, and determination.

4.3 Applications of Emotion Models

Concerning applications, teaching systems have been developed to deal with emotions and thus increase the degree of perseverance and commitment of the students. In parallel, simulators, video games, and ambient-intelligence systems have been developed (see e.g., [Adam et al, 2011] for an overview of the literature and applications of emotions in that domain). Among the very large variety of existing ACA, EM[11] is a typical system that simulates the decline of emotions with time for a specific set of emotions corresponding to the goals that generated them. Another example is the system Affective Reasoner of Gratch & Marsella where agents use representations of themselves and others. Finally, GRETA [de Rosis et al, 2003] is an ACA 3D that can be animated in real time and is capable of expressing emotional states.

5 Conclusion

In the present chapter, we have first tackled the formalization of cognitive-agent systems. Such an agent is capable of behaving in an autonomous way, according to its goals. In addition, it is characterized, a minima, by mental attitudes (beliefs, desires, norms, etc.), time, and action. After a brief overview of the great research avenue in this domain, we have presented the fundamental concept of BDI systems, as well as the tools to deal with the well-known AI problem of knowledge evolution. Finally, we used the aforementioned material in order to formalize two concepts used in those systems: trust and emotion. We also showed that those two concepts can also be formalized in

[11] it is a system based on the Tok architecture of the project Oz. See http://www.cs.cmu.edu/ afs/cs.cmu.edu/project/oz/web/.

a more numerical fashion, which is less fine from the point of view of the definitions of the concepts of interest, but easier to be applied in concrete frameworks.

Of course, there are many other branches in AI about the formalization of cognitive-agent systems. But, some of them are not based on mental states, other are limited to a certain formal language. The peculiarity of the systems presented in this chapter is that they correspond to logic (with both a semantics and an axiomatization) whose properties (in terms of complexity, decidability, and completeness) are also studied. More precisely, those logics are modal logics particularly convenient to represent mental states, as well as relations between those states (beliefs about beliefs, goals, etc. of other agents. The objective is to represent in a fine grain the concepts used by the agents with a logic having "good" logical properties. So, the issues are both computational and mathematical. In addition, they are strongly related to SHS via philosophy and psychology, in particular. It is worth noting that there are studies about the influence of trust on emotions, and vice versa (see e.g., [Bonnefon et al, 2009]).

Naturally, trust and emotion are not the only concepts investigated in the literature. In particular, we have not presented non-reductionist social concepts, for example, notions of group belief or acceptance that are reducible to the sum, over all agents of the group, of their beliefs or acceptance. Consequently, it is necessary to capture a group as a unique entity constituting an institution ruled by specific social rules.

The study of formal properties of intelligent agents is thus a first step in the study of multi-agent systems. The latter need to capture the nature of the group constituted by the agents (What unites them? What is the structure of the group represented by them? Is it just a set of agents or a more complex relational structure including e.g. friendship, hierarchy, commerce, etc.?).

References

Adam C, Herzig A, Longin D (2009) A logical formalization of the OCC theory of emotions. Synthese 168(2):201–248, URL `\url{ftp://ftp.irit.fr/IRIT/LILAC/Journaux_internationaux/2009_Adam_et_al_Synthese.pdf}`

Adam C, Gaudou B, Longin D, Lorini E (2011) Logical modeling of emotions for Ambient Intelligence. In: Mastrogiacomo F, Chong NY (eds) Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives, IGI Global

Amgoud L, Rahwan I (2006) An Argumentation-based Approach for Practical Reasoning. In: Weiss G, Stone P (eds) Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), ACM Press, pp 347–354

Austin JL (1962) How To Do Things With Words. Oxford University Press
Baltag A, Moss LS (2004) Logics for epistemic programs. Synthese 139(2):165–224
Belnap N, Perloff M, Xu M (2001) Facing the future: agents and choices in our indeterminist world. Oxford University Press, New York
van Benthem J, Liu F (2007) Dynamic logic of preference upgrade. Journal of Applied Non-Classical Logics 17(2):157–182
Berreby F, Bourgne G, J-G G (2015) Modelling moral reasoning and ethical responsibility with logical programming. In: Logic for Programming, Artificial Intelligence, and Reasoning, LNCS, vol 9450, Springer, pp 532–548
Bonatti PA, Oliveira EC, Sabater-Mir J, Sierra C, Toni F (2014) On the integration of trust with negotiation, argumentation and semantics. Knowledge Eng Review 29(1):31–50, DOI 10.1017/S0269888913000064, URL https://doi.org/10.1017/S0269888913000064
Bonnefon JF, Longin D, Nguyen MH (2009) A Logical Framework for Trust-Related Emotions. Electronic Communications of the EASST, Formal Methods for Interactive Systems 2009 22:1–16
Bratman M (1987) Intentions, plans, and practical reason. Harvard University Press, Cambridge
Burgess JP (2002) Basic tense logic. In: Gabbay D, Guenthner F (eds) Handbook of Philosophical Logic, vol 7, 2nd edn, Kluwer, pp 1–42
Castaneda HN (1975) Thinking and Doing. D. Reidel, Dordrecht
Castelfranchi C, Paglieri F (2007) The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions. Synthese 155:237–263
Castelfranchi C, Tan YH (eds) (2001) Trust and Deception in Virtual Societies. Kluwer Academic Publishers, Dordrecht
Chellas BF (1980) Modal Logic: an Introduction. Cambridge
Cohen PR, Levesque HJ (1990) Intention is choice with commitment. Artificial Intelligence Journal 42(2–3):213–261
Cohen PR, Morgan J, Pollack ME (eds) (1990) Intentions in Communication. MIT Press, Cambridge, MA
Conte R, Castelfranchi C (1995) Cognitive and social action. London University College of London Press, London
van Ditmarsch H, van der Hoek W, Kooi B (2007) Dynamic Epistemic Logic. Kluwer Academic Publishers
Ditmarsch Hv, der Hoek Wv, Kooi B (2007) Dynamic Epistemic Logic. Kluwer Academic Publishers
Dubois D, Lorini E, Prade H (2017) The strength of desires: A logical approach. Minds and Machines 27(1):199–231, DOI 10.1007/s11023-017-9426-5, URL https://doi.org/10.1007/s11023-017-9426-5
El-Nasr MS, Yen J, Ioerger TR (2000) FLAME: Fuzzy logic adaptive model of emotions. Autonomous Agents and Multi-Agent Systems 3(3):219–257

Gabbay D, Horty J, Parent X, van der Meyden R, van der Torre L (eds) (2013) Handbook of Deontic Logic and Normative Systems. College Publication, http://www.collegepublications.co.uk/downloads/handbooks00001.pdf

Gilbert M (1989) On Social Facts. Routledge, London and New York

Gochet P, Gribomont P (2006) Epistemic Logic. In: Gabbay D, Woods J (eds) Handbook of the History of Logic, vol 7, Elsevier, pp 99–195

Gordon R (1987) The structure of emotions. Cambridge University Press, New York

Gratch J, Marsella S (2005) Lessons from emotion psychology for the design of lifelike characters. Journal of Applied Artificial Intelligence (special issue on Educational Agents - Beyond Virtual Tutors) 19(3-4):215–233

Harel D, Kozen D, Tiuryn J (2000) Dynamic Logic. MIT Press, Cambridge

Herzig A, Lorini E, Hübner JF, Vercouter L (2010) A logic of trust and reputation. Logic Journal of the IGPL 18(1):214–244

van der Hoek W, Jamroga W, Wooldridge M (2007) Towards a theory of intention revision. Synthese 155(2):265–290

Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The Eigentrust Algorithm for Reputation Management in P2P Networks. In: 12th International Conference on World Wide Web (WWW), ACM, pp 640–651

Kooi B (2007) Expressivity and completeness for public update logic via reduction axioms. Journal of Applied Non-Classical Logics 17(2):231–253

Koster A, Schorlemmer WM, Sabater-Mir J (2013) Opening the black box of trust: reasoning about trust models in a BDI agent. J Log Comput 23(1):25–58, DOI 10.1093/logcom/exs003, URL `https://doi.org/10.1093/logcom/exs003`

Lane R, Nadel L (eds) (2000) The cognitive neuroscience of emotions. Oxford

Langville AN, Meyer CD (2005) Deeper Inside PageRank. Internet Mathematics 1(3):335–400

Laverny N, Lang J (2005) From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In: Proceedings of IJCAI'05, Professional Book Center, pp 497–502

Lazarus RS (1991) Emotion and Adaptation. Oxford University Press

van Linder B, van der Hoek, W JJC Meyer (1998) Formalising abilities and opportunities. Fundamenta Informaticae 34:53–101

Loewenstein G (2000) Emotions in economic theory and economic behavior. American Economic Review 90(2):426–432

Lorini E (2011) The cognitive anatomy and functions of expectations revisited. In: Paglieri F, Tummolini L, Falcone R, Miceli M (eds) The Goals of Cognition: Festschfit for Cristiano Castelfranchi, College Publications, London, to appear

Lorini E (2016) A logic for reasoning about moral agents. Logique & Analyse 58(230):177–218

Lorini E, Schwarzentruber F (2011) A logic for reasoning about counterfactual emotions. Artificial Intelligence 175:814–847

Lorini E, Longin D, Gaudou B, Herzig A (2009) The logic of acceptance: grounding institutions on agents' attitudes. Journal of Logic and Computation 19(6):901–940, URL \url{ftp://ftp.irit.fr/IRIT/LILAC/JLC.pdf}

Marsh S (1994) Formalising Trust as a Computational Concept. PhD thesis, Department of Computing Sciece and Mathematics, University of Sterling

Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge, MA

Osman N, Gutierrez P, Sierra C (2015) Trustworthy advice. Knowl-Based Syst 82:41–59, DOI 10.1016/j.knosys.2015.02.024, URL https://doi.org/10.1016/j.knosys.2015.02.024

Page L, Brin S, Motwani R, Winograd T (1998) The PageRank Citation Ranking: Bringing Order to the Web. Tech. rep., Stanford Digital Library Technologies Project

Pelachaud C (2009) Modelling multimodal expression of emotion in a virtual agent. Philosophical transactions of the Royal society B 364:3539–3548

Pinyol I, Sabater-Mir J (2013) Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev 40(1):1–25, DOI 10.1007/s10462-011-9277-z, URL https://doi.org/10.1007/s10462-011-9277-z

Rao AS, Georgeff MP (1991) Modeling rational agents within a BDI-architecture. In: Proceedings of KR'91, Morgan Kaufmann Publishers, pp 473–484

de Rosis F, Pelachaud C, Poggi I, Carofiglio V, De Carolis B (2003) From greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. International Journal of Human-Computer Studies 59:81–118

Sabater J, Sierra C (2005) Review on Computational Trust and Reputation Models. Artificial Intelligence 24:33–60

Searle JR (1969) Speech acts: An essay in the philosophy of language. Cambridge

Searle JR (1983) Intentionality: An essay in the philosophy of mind. Cambridge

Segerberg K (1992) Getting started: Beginnings in the logic of action. Studia Logica 51(3-4):347–378

Segerberg K (1995) Belief revision from the point of view of doxastic logic. Logic Journal of IGPL 3(4):535–553

Singh MP (1999) An ontology for commitments in multiagent systems. Artificial Intelligence and Law 7:97–113

Steunebrink BR, Dastani M, Meyer JJC (2008) A formal model of emotions: integrating qualitative and quantitative aspects. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008), IOS Press, pp 256–260

Turrini P, Meyer JJC, Castelfranchi C (2010) Coping with shame and sense of guilt: a dynamic logic account. Journal of AAMAS 20(3)

van Benthem J (1991) The Logic of Time. D. Reidel Publishing Company
Wooldridge M (2000) Reasoning about Rational Agents. MIT Press
Zeelenberg M, van Dijk WW, Manstead ASR (1998) Reconsidering the relation between regret and responsibility. Organizational Behavior and Human Decision Processes 74:254–272