



**HAL**  
open science

## Logics for games, emotions and institutions

Emiliano Lorini

► **To cite this version:**

Emiliano Lorini. Logics for games, emotions and institutions. *IfColog Journal of Logics and their Applications (FLAP)*, 2017, 4 (9), pp.3075-3113. hal-02640857

**HAL Id: hal-02640857**

**<https://hal.science/hal-02640857>**

Submitted on 28 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Open Archive Toulouse Archive Ouverte


OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible

This is an author's version published in: <http://oatao.univ-toulouse.fr/22057>

### Official URL:

<https://dblp.org/rec/journals/flap/Lorini17>

### To cite this version:

Lorini, Emiliano  *Logics for games, emotions and institutions*. (2017) The IfCoLog Journal of Logics and their Applications, 4 (9). 3075-3113. ISSN 2055-3706.

Any correspondence concerning this service should be sent to the repository administrator: [tech-oatao@listes-diff.inp-toulouse.fr](mailto:tech-oatao@listes-diff.inp-toulouse.fr)

# LOGICS FOR GAMES, EMOTIONS AND INSTITUTIONS

EMILIANO LORINI

*IRIT-CNRS, Toulouse University, France*

`Emiliano.Lorini@irit.fr`

## **Abstract**

We give an informal overview of the way logic and game theory have been used in the past and are currently used to model cognitive agents and multi-agent systems (MAS). In the first part of the paper we consider formal models of mental attitudes and emotions, while in the second part we move from mental attitudes to institutions via collective attitudes.

## **1 Introduction**

Agents in the societies can be either human agents or artificial agents. The focus of this paper is both on: (i) the present society in which human agents interact with the support of ICT through social networks and media, and (ii) the future society with mixed interactions between human agents and artificial systems such as autonomous agents and robots. Indeed, new technologies will come for future society in which such artificial systems will play a major role, so that humans will necessarily interact with them in their daily lives. This includes autonomous cars and other vehicles, robotic assistants for rehabilitation and for the elderly, robotic companions for learning support.

There are two main general observations underlying the present paper. The first is that interaction plays a fundamental role in existing information and communication technologies (ICT) and applications (e.g., Facebook, Ebay, peer-to-peer systems) and will become even more fundamental in future ICT. The second is that the cognitive aspect is crucial for the design of intelligent systems that are expected to interact with human agents (e.g., embodied conversational agents, robotic assistants, etc.). The system must be endowed with a psychologically plausible model of reasoning and cognition in order to be able (i) to understand the human agent's needs and to predict her behaviour, and (ii) to behave in a believable way thereby meeting the human agent's expectations.

Formal methods have been widely used in artificial intelligence (AI) and in the area of multi-agent systems (MAS) for modelling intelligent systems as well as different aspects of social interaction between artificial and/or human agents. The aim of the present paper is to

offer a general overview of the way logic and game theory have been and can be used in AI in order to build formal models of socio-cognitive, normative and institutional phenomena.

We take a bottom-up perspective to the analysis of normative and institutional facts that is in line with some classical analysis in organization theory such as the one presented in March & Simon's famous book "Organizations" [102], described as a book in which they:

"...surveyed the literature on organization theory, starting with those theories that viewed the employee as an instrument and physiological automaton, proceeding through theories that were centrally concerned with the motivational and affective aspects of human behavior, and concluding with theories that placed particular emphasis on cognitive processes" [102, p. 5].

The present paper is organized in two main sections. Section 2 is devoted to cognitive aspects, while Section 3 is devoted to institutional ones. Section 2 starts from the assumption that cognitive agents are, by definition, endowed with a variety of mental attitudes such as beliefs, desires, preferences and intentions that provide input for practical reasoning and decision-making, trigger action execution, and generate emotional responses. We first present a conceptual framework that:

- clarifies the relationship between intention and action and the role of intention in practical reasoning;
- explains how moral attitudes such as standards, ideals and moral values influence decision-making;
- explains how preferences are formed on the basis of desires and moral values;
- clarifies the distinction between the concept of goal and the concept of preference;
- elucidates how mental attitudes including beliefs, desires and intentions trigger emotional responses, and how emotions retroactively influence decision-making and mental attitudes by triggering belief revision, desire change and intention reconsideration.

Then, we explain how game theory and logic have been used in order to develop formal models of such cognitive phenomena. We put special emphasis on a specific branch of game theory, called epistemic game theory, and on a specific family of logics, so-called agent logics. The aim of epistemic game theory is to extend the classical game-theoretic framework with mental notions such as the concepts of belief and knowledge, while agent logics are devoted to explain how different types of mental attitudes (e.g., belief, desires, intentions) are related, how they influence decision and action, and how they trigger emotional responses.

Section 3 builds the connection between mental attitudes and institutions passing by the concept of collective attitude. Collectives attitudes such as joint intention, group belief, group goal, collective acceptance and joint commitment have been widely explored in the area of collective intentionality, the domain of social philosophy that studies how agents function and act at the group level and how institutional facts relate with physical (brute) facts (cf. [100; 140] for a general introduction of the research in this area). Section 3 is devoted to explain (i) how collective attitudes such as collective acceptance or common belief are formed either through aggregation of individual attitudes or through a process of joint perception, (ii) how institutional facts are grounded on collective attitudes and, in particular, how the existence and modification of institutional facts depend on the collective acceptance of these facts by the agent in the society and on the evolution of this collective acceptance. We also discuss existing logics for institutions that formalize the connection between collective attitudes and institutional facts.

In Section 4 we conclude by briefly considering the opposite path leading from norms and institutions to minds. In particular, we explain how institutions and norms, whose existence depends on their acceptance by the agents in the society, retroactively influence the agents' mental attitudes, decisions and actions.

## 2 Mental attitudes and emotions

In this section, we start with a discussion of two issues related with the representation of mental attitudes and emotions: (i) the cognitive processing leading from goal generation to action (Section 2.1), and (ii) the representation of the cognitive structure of emotions and of their influence on behaviour (Section 2.2). Then, we briefly explain how these cognitive aspects have been incorporated into game theory (Section 2.3). Finally, we consider how mental attitudes and emotion are formalized in logic and the connection between the representation of mental attitudes in logic and the representation of mental attitudes in game theory (Section 2.4).

### 2.1 A cognitive architecture

The conceptual background underlying our view of mental attitudes is summarized in Figure 1. (Cf. [90] for a logical formalization of some aspects of this view.) The cognitive architecture represents the process leading from generation of desires and moral values and formation of beliefs via sensing to action performance.

**The origin of beliefs, desires and moral values** An important and general distinction in philosophy of mind is between epistemic attitudes and motivational attitudes. This dis-

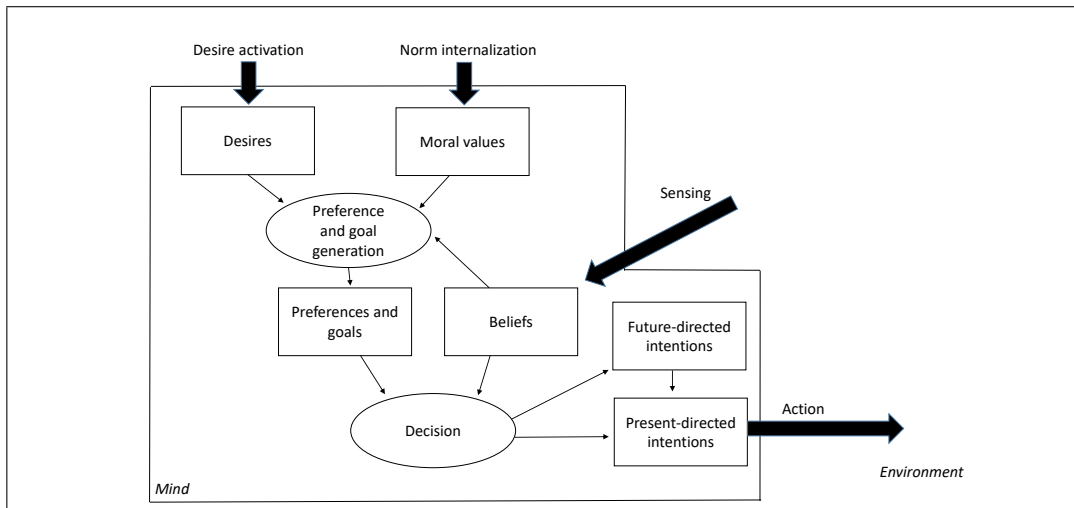


Figure 1: Cognitive architecture

tion is in terms of the *direction of fit* of mental attitudes to the world. While epistemic attitudes aim at being true and their being true is their fitting the world, motivational attitudes aim at realization and their realization is the world fitting them [114; 7; 67]. Searle [125] calls “mind-to-world” the first kind of *direction of fit* and “world-to-mind” the second one.

There are different kinds of epistemic and motivational attitudes with different functions and properties. Examples of epistemic attitudes are beliefs, knowledge and opinions, while examples of motivational attitudes are desires, preferences, moral values and intentions. However, the most primitive and basic forms of epistemic and motivational attitudes are beliefs, desires and moral values.

Beliefs are mental representations aimed at representing how the physical, mental and social worlds are. Indeed, there are beliefs about natural facts and physical events (e.g., I believe that tomorrow will be a sunny day), introspective beliefs (e.g., I believe that I strongly wish that tomorrow will be a sunny day), and beliefs about mental attitudes of other agents (e.g., I believe that you believe that tomorrow will be a sunny day).

Following the Humean conception, a desire can be viewed as an agent’s attitude consisting in an anticipatory mental representation of a pleasant state of affairs (representational dimension of desires) that motivates the agent to achieve it (motivational dimension of desires). The motivational dimension of an agent’s desire is realized through its representational dimension, in the sense that, a desire motivates an agent to achieve it *because* the agent’s anticipatory representation of the desire’s content gives her pleasure so that the agent is “attracted” by it. For example when an agent desires to eat sushi, she is pleased to

imagine herself eating sushi. This pleasant representation motivates her to go to the “The Japoyaki” restaurant in order to eat sushi. This view of desires unifies the standard theory of desire (STD) — focused on the motivational dimension — and the hedonic theory of desire (HTD) — focused on the hedonic dimension —. A third theory of desire has been advanced in the philosophical literature (see [124]), the so-called reward theory of desire (RTD). According to RTD what qualifies a mental attitude as a desire is the exercise of a capacity to represent a certain fact as a reward.<sup>1</sup>

Another fundamental aspect of desire is the *longing aspect*. The idea is that for an agent to desire something, the agent should be in a situation in which she does not have what she desires and she yearns for it. In other words, a state of affairs is desired by an agent only if the agent conceives it as *absent*. The following quotation from Locke [86, Book II, Chap. XXI] makes this point clear:

To return then to the inquiry, what is it that determines the will in regard to our actions? And that...is not, as is generally supposed, the greater good in view: but some (and for the most part the most pressing) uneasiness a man is at present under. This that which successively determines the will, and sets us upon those actions, we perform. This uneasiness we may call, as it is, desire; which is uneasiness of the mind for want of some *absent good*...

This quotation seems in contradiction with what we claimed above, namely, that desire is based on the anticipatory representation of a pleasant state of affairs. However, the stronger the anticipated pleasure associated with a desire, the more painful is its current lack of fulfillment — the term “uneasiness” in the previous quotation —, as in the case of longing for a drink when thirsty, for instance. So the contradiction is only apparent. This aspect of uneasiness described by Locke should not be confused with the concept of aversion which is traditionally opposed to the concept of desire (see [124, Chap. 5]). As emphasized above, if an agent desires a certain fact to be true, then she possesses an anticipatory mental representation of a *pleasant* fact motivating her to make the fact true. On the contrary, if an agent is averse to something, then she possesses an anticipatory mental representation of an *unpleasant* fact motivating her to prevent the fact from being true.

Moral values, and more generally moral attitudes (ideals, standards, etc.), originate from an agent’s capability of discerning what from her point of view is (morally) good from what is (morally) bad. If an agent has a certain ideal  $\varphi$ , then she thinks that the realization of the state of affairs  $\varphi$  ought to be promoted because  $\varphi$  is good in itself. Differently from desires, moral values do not necessarily have a hedonic and somatic component: their

---

<sup>1</sup>According to [39], desire is also a necessary condition for reward. In particular, desire determines what counts as a reward for an agent. For example, a person can be rewarded with water only if she is thirsty and she desires to drink.

fulfillment does not necessarily give pleasure and their transgression does not necessarily give displeasure ‘felt’ from the body.

There are different ways to explain the origin of beliefs, desires, moral values. Beliefs are formed either via direct sensing from the external environment (e.g., I believe that there is a fire in the house since I can see it), communication (e.g., I believe that there is a fire in the house since you told me this and I trust what you say) and inference (e.g., I believe that there is a fire in the house since I already believe that smoke comes out from the house and if there is smoke coming out from the house then there is fire). One might argue that belief formation via direct sensing is more primitive than belief formation via communication and that the latter can be reduced to the former. Indeed, in the context of communication, the hearer first *perceives* the speaker’s utterance, which is nothing but the performance of a physical action (e.g., uttering a certain sound, performing a certain gesture, emitting a certain light signal, etc.) and forms a belief about what the speaker has uttered. Then, she infers the meaning of the speaker’s utterance (i.e., what the speaker wants to express by uttering a certain sound, by performing a certain gesture, by emitting a certain light signal, etc.). Although this is true for communication between humans and between artificial systems situated in the physical environment such as robots, it is not necessarily true for communication in an artificial domain in which there is no precise distinction between an utterance and its meaning. In the latter situation, the speaker may transmit to the hearer a message (e.g., a propositional formula) with a precise and non-ambiguous meaning.

The concept of trust plays a fundamental role in belief formation via direct sensing and via communication. Indeed, the hearer will not believe what the speaker says unless she believes that the speaker is a reliable source of information, thereby trusting the speaker’s judgment. Similarly, for belief formation via direct sensing, an agent will not believe what she sees unless she believes that her perceptual apparatus works properly, thereby trusting it. The issue whether trust is reducible to other mental attitudes is relevant here. A justifiable approach consists in conceiving *communication-based trust* as a belief about the reliability of a source of information, where “reliable” means that, in the normal conditions, what the source says about a given issue is true.

The explanation about the origin of desires adopted in Figure 1 is that they are activated under certain conditions. For instance, according to Maslow’s seminal theory of human motivation, “...everyday conscious desires are to be regarded as symptoms, as surface indicators of more basic needs” [103, p. 392]. Maslow identified a set of basic (most of the time unconscious) needs of human agents including physiological needs,<sup>2</sup> need for safety, need for love and belonging, need for self-esteem and need for self-actualization. For example, a human agent’s desire of drinking a glass of water could be activated by her basic

---

<sup>2</sup>Maslow referred to the concept of homeostasis, as the living system’s automatic efforts to maintain a constant, normal state of the blood stream, body temperature, and so on.



physiological need for bodily balance including a constant body temperature, constant salt levels in the body, and so on. If certain variables of the agent's body are unbalanced and this unbalance is detected,<sup>3</sup> the agent receives a negative unpleasant signal from her body thereby entering in a state of felt displeasure and uneasiness — in the Lockean sense —. Consequently, she becomes intrinsically motivated to restore bodily balance. The connection between the agent's basic need for bodily balance and the agent's desire of drinking a glass of water may rely on the agent's previous experiences and be the product of operant conditioning (also called instrumental learning). Specifically, the agent may have learnt that, under certain conditions, drinking a glass of water is “a suitable means for” restoring balance of certain variables of the body. Indeed, every time the agent drank water when she was feeling thirsty, she got a reward by making her basic need for bodily balance satisfied.<sup>4</sup>

In the case of artificial agents, conditions of desire activation should be specified by the system's designer. For example, a robotic assistant who has to take care of an old person could be designed in such a way that, every day at 4 pm, the desire of giving a medicine to the old person is activated in its mind.

As for the origin of moral values, social scientists (e.g., [6]) have defended the idea that there exist innate moral principles in humans such as fairness which are the product of biological evolution. Other moral values, as highlighted in Figure 1, have a cultural and social origin, as they are the product of the internalization of some external norm. A possible explanation is based on the hypothesis that moral judgments are true or false only in relation to and with reference to one or another agreement between people forming a group or a community. More precisely, an agent's moral values are simply norms of the group or community to which the agent belongs that have been internalized by the agent. This is the essence of the philosophical doctrine of moral relativism (see, e.g., [20]). For example, suppose that an agent believes that in a certain group or community there exists a norm (e.g., an obligation) prescribing that a given state of affairs should be true. Moreover, assume that the agent identifies herself as a member of this group or community. In this case, the agent will internalize the norm, that is, the external norm will become a moral value of the agent and will affect the agent's decisions. For example, suppose that a certain person is (and identifies herself as) citizen of a given country. As in every civil country, it is prescribed that citizens should pay taxes. Her sense of national identity will lead the person to adopt the obligation by imposing the imperative to pay taxes to herself. When deciding to pay taxes or not, she will decide to do it, not simply in order to avoid being sanctioned and being exposed to punishment, but also because she is motivated by the moral obligation

---

<sup>3</sup>Converging empirical evidences from neuroscience show that the hypothalamus is responsible for monitoring these bodily conditions.

<sup>4</sup>Following [124], one might argue that most conscious desires (including the desire to eat at a particular time and the desire to drink water) are instrumental, as they are activated *in order to* satisfy more basic needs of the individual.

to paying taxes.

**From desires and moral values to preferences** According to contemporary theories of human motivation both in philosophy and in economics (e.g., [127; 60]), preferences of a rational agent may originate either (i) from somatically-marked motivations such as desires or physiological needs and drives (e.g., the goal of drinking a glass of water originated from the physiological drive of thirst), or (ii) from moral considerations and values (e.g., the goal of helping a poor person originated from the moral value of taking care of needy people). More generally, there exists desire-dependent preferences and desire-independent ones originated from moral values. This distinction allows us to identify two different kinds of moral dilemmas. The first kind of moral dilemma is the one which is determined by the logical conflict between two moral values. The paradigmatic example is the situation of a soldier during a war. As a member of the army, the soldier feels obliged to kill his enemies, if this is the only way to defend his country. But, as a catholic, he thinks that human life should be respected. Therefore, he feels morally obliged not to kill other people. The other kind of moral dilemma is the one which is determined by the logical conflict between desires and moral values. The paradigmatic example is that of Adam and Eve in the garden of Eden. They are tempted by the desire to eat the forbidden fruit and, at the same time, they have a moral obligation not to do it.

According to the cognitive architecture represented in Figure 1, desires and moral attitudes of an agent are two different parameters affecting the agent's preferences. This allows us to draw the distinction between *hedonistic* agents and *moral* agents. A purely hedonistic agent is an agent who acts in order to maximize the satisfaction of her own desires, while a purely moral agent is an agent who acts in order to maximize the fulfillment of her own moral values. In other words, if an agent is purely hedonistic, the utility of an action for her coincides with the personal good the agent will obtain by performing this action, where the agent's personal good coincides with the satisfaction of the agent's own desires. If an agent is purely moral, the utility of an action for her coincides with the moral good the agent will promote by performing this action, where the agent's promotion of the moral good coincides with the accomplishment of her own moral values. Utility is just the quantitative counterpart of the concept of preference, that is, the more an agent prefers something, the higher its utility. Of course, purely hedonistic agents and purely moral agents are just extremes cases. An agent is more or less moral depending on whether the utility of a given option for her is more or less affected by her moral values. More precisely, the higher is the influence of the agent's moral values on evaluating the utility of a given decision option, the more moral the agent is. The extent to which an agent's utility is affected by her moral values can be called *degree of moral sensitivity*.<sup>5</sup>

---

<sup>5</sup>This degree can be conceived as a personality trait. In the case of human agents, it is either culturally

**Goals** The reason why, in Figure 1, preferences and goals are included in the same box is that we conceive goals as intimately related with preferences. In particular, we assume that an agent has  $\varphi$  as a goal (or wants to achieve  $\varphi$ ) if and only if: (i) the agent prefers  $\varphi$  to be true to  $\varphi$  to be false, and (ii) the agent considers  $\varphi$  a possible state of affairs ( $\varphi$  is compatible with what the agent believes). The second property is called *realism* of goals by philosophers (cf. [22; 37; 104]). It is based on the idea that an agent cannot reasonably pursue a goal unless she thinks that she can *possibly* achieve it, i.e., there exists at least one possible evolution of the world (a history) that the agent considers possible along which  $\varphi$  is true. Indeed, an agent's goal should not be incompatible with the agent's beliefs. This explains the influence of beliefs on the goal generation process, as depicted in Figure 1.<sup>6</sup> The first property is about the motivational aspect of goals. For  $\varphi$  to be a goal, the agent should not be indifferent between  $\varphi$  and  $\neg\varphi$ , in the sense that, the agent prefers a situation in which  $\varphi$  is true to a situation in which  $\varphi$  is false, *all other things being equal*. In other words, the utility of a situation increases in the direction by the formula  $\varphi$  *ceteris paribus* (“all else being equal”) [154]. This property also defines Von Wright's concept of “preference of  $\varphi$  over  $\neg\varphi$ ” [152].<sup>7</sup> According to this interpretation, a goal is conceived as a *realistic ceteris paribus preference for  $\varphi$* .

However not all goals have the same status. Certain goals have a motivating force while others do not have it. Indeed, the fact that the agent prefers  $\varphi$  being true to  $\varphi$  being false does not necessarily imply that the agent is motivated to achieve a state in which  $\varphi$  is true and that she decides to perform a certain action *in order to* achieve it. For  $\varphi$  to be a motivating goal, for every possible situation that the agent envisages in which  $\varphi$  is true and for every possible situation that the agent envisages in which  $\varphi$  is false, the agent has to prefer the former to the latter. In other words, there is no way for the agent to be satisfied without achieving  $\varphi$ .<sup>8</sup>

An example better clarifies this point. Suppose Mary wants to buy a reflex camera Nikon and, at the same time, she would like to spend no more than 300 euros. In other words, Mary has two goals in her mind:

- G1: the goal of buying a reflex camera Nikon, and
- G2: the goal of spending no more than 300 euros.

She goes to the shop and it turns out that all reflex cameras Nikon cost more than 300 euros. This implies that Mary believes that she cannot achieve the two goals at the same

---

acquired or genetically determined. In the case of artificial agents, it is configured by the system designer.

<sup>6</sup>The idea that beliefs form an essential ingredient of the goal generation process is also suggested by [26].

<sup>7</sup>Von Wright presents a more general concept of “preference of  $\varphi$  over  $\psi$ ” which has been recently formalized in a modal logic setting by [146]. See also [119] for an interpretation of this *ceteris paribus* condition based on the concept of logical independence between formulas.

<sup>8</sup>The term ‘satisfied’ just means that the agent achieves what she prefers.

time, as she envisages four situations in her mind but only three are considered possible by her: the situation in which only the goal G1 is achieved, the situation in which only the goal G2 is achieved and the situation in which no goal is achieved. The situation in which both goals are achieved is considered impossible by Mary. This is not inconsistent with the previous definition of goal since Mary still believes that it is possible to achieve each goal separately from the other. Figure 2 clearly illustrates this: the full rectangle includes all worlds that Mary envisages, so-called *information set*, while the dotted rectangle includes all worlds that Mary considers actually possible, so-called *belief set*.<sup>9</sup> (Cf. [76; 91] for a logical account of the distinction between information set and belief set.)

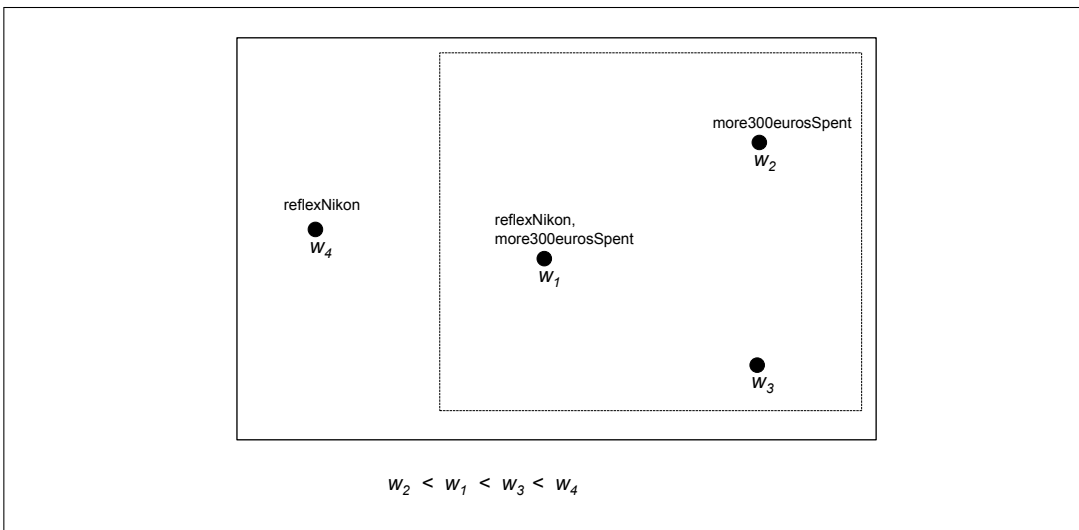


Figure 2: Example for goals

Mary decides to save her money since the goal G2 is a motivating one, while the goal G1 is not. To see that G1 is a goal, it is sufficient to observe that, *all other things being equal*, Mary prefers a situation in which she buys a Nikon to the situation in which she does not buy it. In fact,  $w_4$  is preferred to  $w_3$  and  $w_1$  is preferred to  $w_2$ . Moreover,  $w_4$  and  $w_3$  are equal in everything except at  $w_4$  Mary buys a Nikon while at  $w_3$  she does not. Similarly,  $w_1$  and  $w_2$  are equal in everything except at  $w_1$  Mary buys a Nikon while at  $w_2$  she does not. To see that G1 is not motivating, it is sufficient to observe that there exists a situation in which Mary does not buy a Nikon ( $w_3$ ) that is preferred to a situation in which she does it ( $w_1$ ). Finally, to see that G2 is a motivating goal, we just need to observe that

<sup>9</sup>Mary's information set includes all worlds that, according to Mary, are compatible with the laws of nature. For instance, Mary can perfectly envisage a world in which she is the president of French republic even though she considers this actually impossible.

every situation in which she spends no more than 300 euros ( $w_3$  and  $w_4$ ) is preferred to every situation in which this is not the case ( $w_1$  and  $w_2$ ). Thus, on the basis of what she believes, Mary concludes that she can only achieve her goal G2 by saving her money and by buying nothing in the shop.

**From preferences and beliefs to actions** As the cognitive architecture in Figure 1 highlights, beliefs and preferences are those mental attitudes which determine the agent's choices and are responsible for the formation of new intentions about present actions (present-directed intentions) and future actions (future-directed intentions). As emphasized in the literature in philosophy [22; 105] and AI [23], a future-directed intention is the element of a partial or a complete plan of the agent: an agent may have the intention to perform a sequence of actions later (e.g., the action of going to the train station in two hours followed by the action of taking the train from Paris to Bruxelles at 10 am) in order to achieve a certain goal (e.g., the goal of being in Bruxelles at the European Commission at 2 pm). A present-directed intention is a direct motivation to perform an action now.

In particular, decision is determined by beliefs, preferences and a general rationality criterion stating what an agent should do on the basis of what she believes and what she prefers. Different kinds of rationality criteria have been studied in the areas of decision theory and game theory ranging from expected utility maximization, maxmin and maxmax to satisficing [133]. Once the choice has been made by the agent and the corresponding intention has been formed, the action is performed right afterwards or later. Specifically, an agent forms the intention to perform a certain action at a given point in time and, once the time of the planned action execution is attained, the agent performs the action unless before attaining it, she has reconsidered her prior intention.

## 2.2 A cognitive view of emotion

In the recent years, emotion has become a central topic in AI. The main motivation of this line of research lies in the possibility of developing computational and formal models of artificial agents who are expected to interact with humans. To ensure the accuracy of a such formal models, it is important to consider how emotions have been defined in the psychological literature. Indeed, in order to build artificial agents with the capability of recognizing the emotions of a human user, of behaving in a believable way, of affecting the user's emotions by the performance of actions directed to her emotions (e.g. actions aimed at reducing the human's stress due to his negative emotions, actions aimed at inducing positive emotions in the human), such agents must be endowed with an adequate model of human emotions.

**Appraisal theory** The most popular psychological theory of emotion in AI is the so-called appraisal theory (cf. [123] for a broad introduction to the developments in appraisal theory). This theory has emphasized the strong relationship between emotion and cognition, by stating that each emotion can be related to specific patterns of evaluations and interpretations of events, situations or objects (appraisal patterns) based on a number of dimensions or criteria called *appraisal variables* (e.g. goal relevance, desirability, likelihood, causal attribution). Appraisal variables are directly related to the mental attitudes of the individual (e.g. beliefs, predictions, desires, goals, intentions). For instance, when prospecting the possibility of winning a lottery and considering ‘I win the lottery’ as a desirable event, an agent might feel an intense hope. When prospecting the possibility of catching a disease and considering ‘I catch a disease’ as an undesirable event, an agent might feel an intense fear.

Most appraisal models of emotions assume that explicit evaluations based on evaluative beliefs (i.e. the belief that a certain event is good or bad, pleasant or unpleasant, dangerous or frustrating) are a necessary constituent of emotional experience. On the other hand, there are some appraisal models mostly promoted by philosophers [126; 50] in which emotions are reduced to specific combinations of beliefs and desires, and in which the link between cognition and emotion is not necessarily mediated by evaluative beliefs. Reizenzein [118] calls *cognitive-evaluative* the former and *cognitive-motivational* the latter kind of models. For example, according to cognitive-motivational models of emotions, a person’s happiness about a certain fact  $\varphi$  can be reduced to the person’s belief that  $\varphi$  obtains and the person’s desire that  $\varphi$  obtains. On the contrary, according to cognitive-evaluative models, a person feels happy about a certain fact  $\varphi$  if she believes that  $\varphi$  obtains and she evaluates  $\varphi$  to be good (desirable) for her. The distinction between cognitive-evaluative models and cognitive-motivational models is reminiscent of the opposition between the Humean view and the anti-Humean view of desire in philosophy of mind. According to the Humean view, belief and desires are distinct mental attitudes that are not reducible one to the other. Moreover, according to this view, there are no necessary connections between beliefs and desires, i.e., beliefs do not necessarily require corresponding desires and, viceversa, desires do not necessarily require corresponding beliefs. On the contrary, the anti-Humean view defends the idea that beliefs and desires are necessarily connected. A specific version of anti-Humeanism is the so-called “Desire-as-Belief Thesis” criticized by the philosopher David Lewis in [80] (see also [81; 57]). In line with cognitive-evaluative models, this thesis states that an agent *desires* something to the extent that she *believes* it to be good.

The popularity of appraisal theory in logic and AI is easily explained by the fact that it perfectly fits with the concepts and level of abstraction of existing logical and computational models of cognitive agents developed in these areas. Especially cognitive-motivational models use folk-psychology concepts such as belief, knowledge, desire and intention that are traditionally used in logic and AI for modelling cognitive agents.

The conceptual background underlying our view of appraisal theory is depicted in Figure 3 which is nothing but the cognitive architecture of Figure 1 extended with an emotion component.

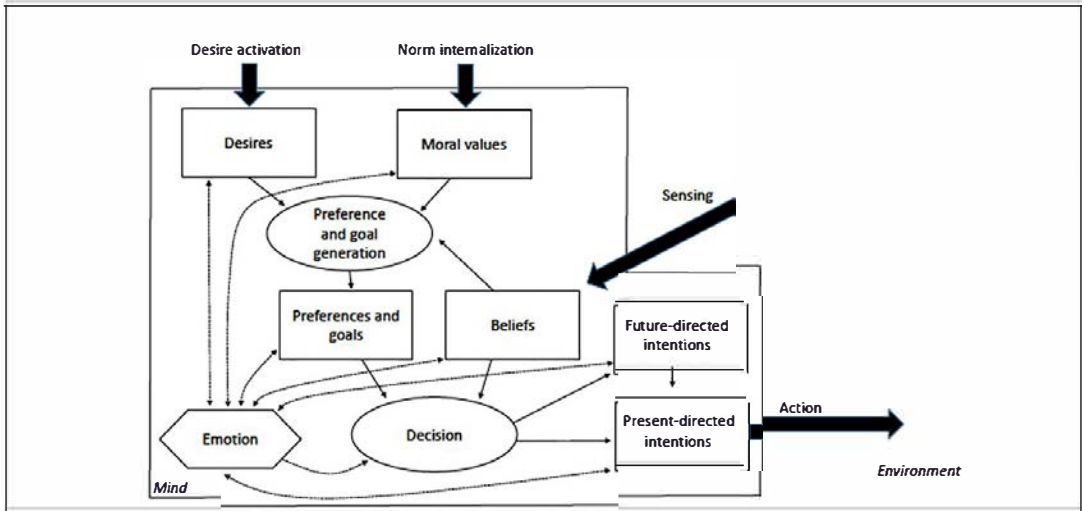


Figure 3: Cognitive architecture extended with emotions

Figure 3 highlights the role of mental attitudes in emotion. In particular, it highlights the fact that mental attitudes of different kinds such as belief, desires, preferences, goals, moral values and (present-directed or future-directed) intentions determine emotional responses. For example, as emphasized above, the emotional response of happiness is triggered by a *goal* and the *certain belief* that the content of one's goal is true. On the contrary, the emotional response of sadness is triggered by a *goal* and the *certain belief* that the content of one's goal is false. The emotional response of hope is triggered by a *goal* and the *uncertain belief* that the content of one's goal is true. On the contrary, the emotional response of fear is triggered by a *goal* and the *uncertain belief* that the content of one's goal is false. This view is consistent with a famous appraisal model, the so-called OCC psychological model of emotions [111], according to which, while joy and distress are triggered by *actual consequences*, hope and fear are triggered by *prospective consequences* (or *prospects*). [52] interpret the term 'prospect' as synonymous of 'uncertain consequence' (in contrast with 'actual consequence' as synonymous of 'certain consequence').

Moral guilt and reproach are examples of emotions that are triggered by moral values [56]. While moral guilt is triggered by the *belief* of being responsible for the violation of a *moral value* or the *belief that one is responsible for having behaved in a morally reprehensible way*, reproach is triggered by the *belief* that someone else is responsible for the violation of a *moral value* or *belief that someone else is responsible for having behaved in a*

*morally reprehensible way*. In other words, guilt is triggered by self-attribution of responsibility for the violation of a moral value, while reproach is triggered by attribution to others of responsibility for the violation of a moral value.

Intentions as well might be responsible for triggering certain kinds of emotional response. For instance, as emphasized by psychological theories of anger (e.g., [77; 111; 121]), a necessary condition for an agent 1 to be angry towards another agent 2 is the agent 1's belief that agent 2 has performed an action that has damaged her, that is, 1 believes that she has been kept from attaining an important goal by an improper action of agent 2. Anger becomes more intense when agent 1 believes that agent 2 has *intentionally* caused the damage. In this sense, an agent 1's belief about another agent 2's intention may have implications on the intensity of agent 1's emotions.

Figure 3 also represents how emotions retroactively influence mental states and decision either (i) through coping or (ii) through anticipation and prospective thinking (i.e., the act of mentally simulating the future) in the decision-making phase.

Coping is the process of dealing with emotion, either externally by forming an intention to act in the world (problem-focused coping) or internally by changing the agent's interpretation of the situation and the mental attitudes that triggered and sustained the emotional response (emotion-focused coping) [77]. For example, when feeling an intense fear due to an unexpected and scaring stimulus, an agent starts to reconsider her beliefs and intentions in order to update her knowledge in the light of the new scaring information and to avoid running into danger (emotion-focused coping). Then, the agent forms an intention to go out of danger (problem-focused coping). Another agent can try to discharge her feeling of guilt for having damaged someone either by forming the intention to repair the damage (problem-focused coping) or by reconsidering the belief about her responsibility for the damage (emotion-focused coping). The coping process as well as its relation with appraisal is illustrated in Figure 4.

**Influence of emotion on decision** The influence of emotion on decision-making has been widely studied both in psychology and in economics. Rick & Loewenstein [120] distinguish the following three forms of influence:

- **Immediate emotions:** real emotions experienced at the time of decision-making:
  - **Integral influences:** influences from immediate emotions that arise from contemplating the consequences of the decision itself,
  - **Incidental influences:** influences from immediate emotions that arise from factors unrelated to the decision at hand (e.g., the agent's current mood or chronic dispositional affect);



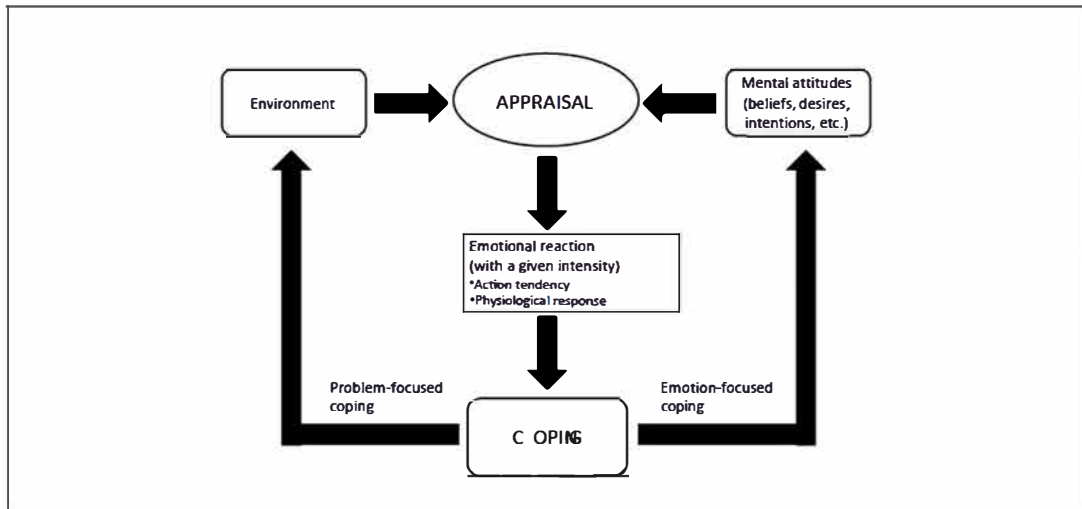


Figure 4: Appraisal and coping cycle

- **Anticipated emotions:** predictions about the emotional consequences of decision outcomes (they are not experienced as emotions per se at the time of decision-making).

An example of integral influence of an immediate emotion is given by the following example.

**Example 1.** *Paul would like to eat some candies but her mother Mary has forbidden him to eat candies without her permission. Paul's fear of the sanction influences Paul's decision not to eat candies without asking permission.*

The following example illustrates incidental influence of an immediate emotion.

**Example 2.** *Mary has quarreled with her colleague Paul. At the end of the day she goes back home after work and on the metro a beggar asks her for money. Few hours after the quarrel with Paul, Mary is still in a bad mood and because of her current disposition she refuses the beggar's request.*

The following example illustrates the influence of anticipated emotions on decision.

**Example 3.** *Peter has to decide whether to leave her job as a researcher at the university of Paris and to accept a job offer as a professor at a university in the U.S. She decides to accept the job offer because she thinks that, if she refuses it, she will likely regret her decision.*

One of the most prominent theory of the integral influence of emotion on decision is Damasio's theory of the somatic marker [35]. According to this theory, decision between different courses of actions leads to potentially advantageous (positive) or harmful (negative) outcomes. These outcomes induce a somatic response used to mark them and to signal their danger or advantage. In particular, a negative somatic marker 'signals' to the agent the fact that a certain course of action should be avoided, while a positive somatic marker provides an incentive to choose a specific course of action. According to Damasio's theory, somatic markers depend on past experiences. Specifically, pain or pleasure experienced as a consequence of an outcome are stored in memory and are felt again when the outcome is envisaged in the decision-making process. The following example clearly illustrates this.<sup>10</sup>

**Example 4.** *Mary lives in Toulouse and has to decide whether to go to Paris by plane or by train. Last time she traveled by plane she had a painful experience because of turbulence. Mary envisages the possibility of incurring again in a turbulence and gets frightened, thereby deciding to travel by train.*

Several works aimed at extending the classical expected utility model to incorporate anticipated emotions that are related to our uncertainty about the future, such as hopefulness, anxiety, and suspense [27]. Some economic models of decision-making consider how the anticipation of a future regret might affect a person's current decision [87]. In particular, according to these models, if a person believes that after choosing a certain action she will likely regret for having made this choice, she will be less willing to choose the action (than in the case in which she does not believe this). These models agree in defining regret as the emotion that stems from the comparison between the actual outcome deriving from a given choice and a counterfactual better outcome that might have been had one chosen a different action [45; 70; 157]. More recently, some economists have studied the influence of strategic emotions such as interpersonal guilt and anger on decision [10; 29; 65]. Following psychological theories of interpersonal guilt [12; 139], models developed in this area assume that the prototypical cause of guilt is the infliction of harm, loss, or distress on a relationship partner. Moreover, they assume that if people feel guilty for hurting their partners and for failing to live up to their expectations, they will alter their behavior (to avoid guilt) in ways that seem likely to maintain and strengthen the relationship. This is different from the concept of moral guilt formalized by [97] according to which a person feels (morally) guilty if she believes that she is responsible for having behaved in a morally reprehensible way (see Section 2.4 for more details).

---

<sup>10</sup>Positive and negative somatic markers can operate either at a conscious level or at a unconscious/automatic level. This corresponds to Ledoux's distinction between explicit memory and implicit memory and between two possible elaborations of a stimulus inducing an emotional response [78]: conscious elaboration vs. automatic elaboration.

### 2.3 Interacting minds: from game theory to epistemic game theory

The idea highlighted in Section 2.1 of describing rational agents in terms of their epistemic and motivational attitudes, is also adopted by classical decision theory and game theory. In particular, classical decision theory accounts for the criteria and principles (e.g., expected utility maximization) that a rational agent should apply in order to decide what to do on the basis of her beliefs and preferences. Game theory generalizes decision theory to the multi-agent case in which agents' decisions are interdependent and agents' actions might interfere between them so that: (i) the possibility for an agent to achieve her goals may depend on what the other agents decide to do, and (ii) agents form beliefs about the future choices of the other players and, consequently, their current decisions are influenced by what they believe the others will do. More generally, game theory involves a strategic component that is not considered by classical decision theory whose object of analysis is a single agent who makes decisions and acts in an environment she does not share with other agents.

Classical decision theory and game theory provide a quantitative account of individual and strategic decision-making by assuming that agents' beliefs and preferences can be respectively modeled by subjective probabilities and utilities. In particular, while subjective probability captures the extent to which a fact is *believed* by a certain agent, utility captures how much a certain state of affairs is *preferred* by the agent. In other words, subjective probability is the quantitative counterpart of the concept of belief, while utility is the quantitative counterpart of the concept of preference.<sup>11</sup>

One of the fundamental concepts of game theory is the concept of solution which is, at the same time, a prescriptive notion, in the sense that it prescribes how rational agents in a given interaction *should* play, and a predictive one, in the sense that it allows us to predict how the agents *will* play. There exist many different solution concepts both for games in normal form and for games in extensive form (e.g., Nash Equilibrium, iterated deletion of strongly dominated strategies, iterated deletion of weakly dominated strategies, correlated equilibrium, backward induction, forward induction, etc.) and new ones have been proposed in the recent years (see, e.g., [58]). A major issue we face when we want to use a solution concept in order either to predict human behavior or to build some practical applications (e.g., for computer security or for multi-agent systems) is to evaluate its significance. Some of the questions that arise in these situations are, for instance: given certain assumptions about the agents such as the assumption that they are rational (e.g., utility

---

<sup>11</sup>Qualitative approaches to individual and strategic decision-making have been proposed in AI [18; 68] to characterize criteria that a rational agent should adopt for making decisions when she cannot build a probability distribution over the set of possible events and her preference over the set of possible outcomes cannot be expressed by a utility function but only by a qualitative ordering over the outcomes. For example, going beyond expected utility maximization, qualitative criteria such as the maxmin principle (choose the action that will minimize potential loss) and the maxmax principle (choose the action that will maximize potential gain) have been studied and axiomatically characterized [19; 20].

maximizers), under which conditions will the agents converge to equilibrium? Are these conditions realistic? Are they too strong for the domain of application under consideration? There is a branch of game theory, called epistemic game theory, which can help to answer these questions (cf. [113] for a general introduction to the research in this area). Indeed, the aim of epistemic game theory is to provide an analysis of the necessary and/or sufficient epistemic conditions of the different solution concepts, that is, the assumptions about the epistemic states of the players that are necessary and/or sufficient to ensure that they will play according to the prescription of the solution concept. Typical epistemic conditions which have been considered are, for example, the assumption that players have common belief (or common knowledge) about the rationality of every player,<sup>12</sup> the assumption that every player knows the choices of the others,<sup>13</sup> or the assumption that players are logically omniscient.<sup>14</sup>

Epistemic game theory shares concepts and methods with what Aumann calls interactive epistemology [8]. The latter is the research area in logic and philosophy which deals with formal models of knowledge and belief when there is more than one rational agent or “player” in the context of interaction having not only knowledge and beliefs about substantive matters, but also knowledge and beliefs about the others’ knowledge and beliefs. The concept of rationality corresponds either to the optimality criterion according to which an agent should choose an action which guarantees the highest utility, given what she believes the other agents will do, or the prudential criterion according to which an agent should not choose an action which ensures the lowest utility, given what she believes the other agents will do. An example of the former is expected utility maximization, while an example of the latter is weak rationality in the sense of [145] (cf. also [109; 15]), according to which an agent should not choose an action which is strongly dominated by another action, given what the agent believes the other agents will do.

Epistemic game theory provides a useful framework for clarifying how agents’ mental attitudes influence behaviours of agents in a social setting. In particular, it allows us to understand the subtle connection between beliefs, preferences and decision, as represented in Figure 1 given in Section 2.1, under the assumption that the agents’ decisions are interdependent, in the sense that they are affected by what the agents believe the others will choose.<sup>15</sup>

---

<sup>12</sup>This is the typical condition of iterated deletion of strongly dominated strategies (also called iterated strong dominance).

<sup>13</sup>This condition is required in order to ensure that the agents will converge to a Nash equilibrium.

<sup>14</sup>See [158] for an analysis of iterated strong dominance after relaxing the assumption of logical omniscience.

<sup>15</sup>Although epistemic game theory and, more generally, game theory share with Figure 1 the concepts of belief and preference, they do not provide an account of the origin of beliefs, desires and moral values and of the connection between desires, moral values and preferences. Moreover, the concept of future-directed intention is not included in the conceptual apparatus of game theory and epistemic game theory. The same can be said

## 2.4 Logics for mental attitudes, emotion and games

This section is devoted to discuss existing logics for mental attitudes and emotion proposed in AI as well as the connection between the representation of mental attitudes and emotion in logic and the representation of mental attitudes and emotion in game theory.

**Logics for mental attitudes** Since the seminal work of [31] aimed at implementing Bratman's philosophical theory of intention [22], many formal logics for reasoning about mental attitudes of agents such as beliefs, desires and intentions have been developed. Among them we should mention the logics developed by [93; 90; 63; 75; 107; 108; 117; 130; 134; 148; 155].

The general term used to refer to this family of logics is *agent logics*. A subfamily is the family of BDI logics whose most representative example is the modal logic by [117] whose primitive constituents are the the concepts of belief (B), desire (D) and intention (I) which are expressed by corresponding modal operators. Another well-known agent logic is the so-called KARO framework developed by [107]. KARO is a multi-modal logic framework based on a blend of dynamic logic with epistemic logic, enriched with modal operators for modeling mental attitudes such as beliefs, desires, wishes, goals and intentions.

Generally speaking, agent logics are nothing but formal models of rational agency whose aim is to explain how an agent endowed with mental attitudes makes decisions on the basis of what she believes and of what she wants or prefers. In this sense, the decisions of the agent are determined by both the agent's beliefs (the agent's epistemic states) and the agent's preferences (the agent's motivational states). As discussed in Section 2.1, the output of the agent's decision-making process is either a choice about what to do in the present, also called present-directed intention, or a choice about what to do in the future, also called future-directed intention. The idea that the behavior of an agent can be explained by attributing mental states to the agent and by having a sophisticated account of the relationship between her epistemic states and her motivational states and of the influence of these on the agent's decision-making process is something agent logics share with other disciplines including philosophy of mind [38], cognitive sciences [116], psychology [118] and artificial intelligence [28].

**Logics for emotion** More recently, agent logics have been used to formalize the cognitive structure and the coping strategies of different types of emotion. For instance, a logical formalization of emotion in the context of the KARO framework has been proposed. In particular, in the KARO framework each emotion type is represented with a special predicate, or fluent, in the jargon of reasoning about action and change, to indicate that these

---

for goals: the concept of goal is somehow implicit in the utility function but is not explicitly modeled.

predicates change over time. For every fluent a set of effects of the corresponding emotions on the agent's planning strategies are specified, as well as the preconditions for triggering the emotion in terms of mental attitudes of agents. The latter correspond to generation rules for emotions. For instance, in [106] generation rules for four basic emotions are given: joy, sadness, anger and fear, depending on the agent's plans. In [144] generation rules for guilt and shame have been proposed.

A logical formalization of the OCC psychological model of emotions [111] has been proposed in [1].

Surprise is the simplest emotion that is triggered by the mismatch between an expectation that an event will possibly occur and an incoming input (i.e., what an agent perceives). In [92] a logical theory of surprise is proposed. The theory clarifies two important aspects of this cognitive phenomenon. First, it addresses the distinction between surprise and astonishment, the latter being the emotion triggered by something an agent could not reasonably expect. The crucial difference between surprise and astonishment is that the former necessarily requires an explicit expectation in the agent's mind, while the latter does not. One can be astonished by something since, at the moment she perceives it, she realizes that it was totally unpredictable, without having formulated an expectation in advance. For example, suppose Mary is working in her office. Suddenly, someone knocks the door and enters into Mary's office. Mary sees that the person is a policeman. She is astonished by this fact even though, before perceiving it, she did not have explicit in her mind the expectation that "a policeman will not enter into the office". Secondly, the theory clarifies the role of surprise in belief change by conceiving it as a basic mechanism which is responsible for triggering belief reconsideration.

In a more recent paper [99], a logical formalization of counterfactual emotions has been provided. Counterfactual emotions, whose prototypical example is regret, are those emotions that are based on counterfactual reasoning about agents' choices. Other examples are rejoicing, disappointment, and elation. The formalization is based on an epistemic extension of STIT logic (the logic of "seeing to it that") by Belnap et al. [13; 66; 25; 89] and allows to capture the cognitive structure of regret and, in particular, the counterfactual belief which is responsible for triggering this emotion, namely the *belief that a counterfactual better outcome might have been, had the agent chosen a different action*. In [96], the STIT logical analysis of counterfactual emotions is extended to moral emotions. The latter involve counterfactual reasoning about responsibility for the transgression of moral values. In particular, the proposed formalization accounts for the attribution of responsibility for the violation of a moral value either to the self or to the other. This is a fundamental constituent of moral emotions such as guilt, reproach, moral pride and moral approval. For example, according to the proposed analysis, guilt is triggered by the *belief that one is responsible for having behaved in a morally reprehensible way*. A game-theoretic account of moral guilt, which parallels the STIT logical analysis, has been given in [97].

The problem of emotion intensity has also been addressed by logicians. Following existing psychological models of emotion based on appraisal theory, intensity of these emotions is defined as a function of two cognitive parameters, the strength of the expectation and the strength of the desire which are responsible for triggering the emotional response. For instance, the intensity of hope that a certain event will occur is a monotonically increasing function of both the strength of the expectation and the strength of the desire that the event will occur. The logical theory of appraisal and coping presented in [36] also considers the behavioral aspects of such emotions: how the execution of a certain coping strategy depends on the intensity of the emotion generating it. Specifically, it is assumed that: (i) an agent is identified with a numerical value which defines her tolerance to the negative emotion, and (ii) if the intensity of the negative emotion (e.g., fear) exceeds this value then the agent will execute a coping strategy aimed at discharging the negative emotion.

**Logics for games** The relationship between logic and game theory has been explored in both directions: *games for logic* and *logic for games*. On the one hand, methods and techniques from game theory have been applied to formal semantics, proof theory and model checking for different kinds of logic [64; 51; 72]. On the other hand, logical representation languages have been proposed in computer science and AI to represent game-theoretic concepts such as the concepts of strategy, capability, winning strategy as well as solution concepts such as Nash equilibrium and backward induction. This includes logics such as Coalition Logic [112], Alternating-time Temporal Logic (ATL) [5] and STIT (the logic of “seeing to it that”) [13; 66].

More recently, logics for epistemic game theory have been proposed by incorporating epistemic components in existing logics for games and developing new logical formalisms that can represent, at the same time, the structure of the game and the mental attitudes and rationality of the players involved in the game.

Much of the work in the field of epistemic game theory is based on a *quantitative* representation of uncertainty and epistemic attitudes. Notable examples are the analysis of the epistemic foundations for forward induction and for iterated admissibility based on Bayesian probabilities [135; 59], conditional probabilities [11] or lexicographic probabilities [21]. The distinction between quantitative and qualitative approaches to uncertainty has been widely discussed in the AI literature (cf. [49]). While in quantitative approaches belief states are characterized by classical probabilistic measures or by alternative numerical accounts, such as lexicographic probabilities or conditional probabilities [11], qualitative approaches do not use any numerical representation of uncertainty but simply a plausibility ordering on possible worlds structures inducing an epistemic-entrenchment-like ordering on propositions.

Both logics for epistemic game theory based on a qualitative representation of epistemic

attitudes [9; 91; 98] and logics for epistemic game theory based on probability theory [59; 17] have been proposed in the recent years. The main motivation for the latter is to exploit logical methods in order to provide sound and complete axiomatics for important concepts studied in epistemic game theory such as rationality and common knowledge of rationality. The main motivation for the former is to show that interesting results about the epistemic foundation for solution concepts in game theory can be proved in a qualitative setting, without necessarily exploiting the complex machinery of probability theory.

The connection between logical models of epistemic states based on Kripke semantics and formal models of epistemic states based on the concept of type space has also been explored [46; 73]. While the former have been mainly proposed by logicians in AI [44] and philosophy [137], the latter have been proposed by game theorists in economics [61]. The main motivation for this research lies in the possibility of building a bridge between two research communities that study the same concepts and phenomena from different perspectives.

### 3 From mental attitudes to institutions via collective attitudes

In this section we gradually move from minds to institutions. The connection between the former and the latter is built via the concept of collective attitude. Specifically, we discuss a particular view of institutions: the idea that institutional facts are grounded on the agents' collective attitudes that, in turn, originate from the agents' mental attitudes.

Section 3.1 starts with a discussion about the different functions and origins of collective attitudes, while Section 3.2 clarifies the connection between collective attitudes and institutions. Finally, Section 3.3 explains how this connection has been formalized in logic.

#### 3.1 Collective attitudes

Collectives such as groups, teams, corporations, organizations, etc. do not have minds. However, we frequently ascribe intentional attitudes to them in the same way as we ascribe intentional attitudes to individuals. For example, we may speak of what our family prefers, of what the goal of a corporation or organization is, of what the scientific community think about a certain issue, and so on.

**Aggregate vs. common attitudes** An important distinction in the theory of collective attitudes is between aggregate attitudes and common attitudes. As emphasized by [84] "...an aggregate attitude (of a collective) is an aggregate or summary of the attitudes of the individual members of the collective, produced by some aggregation rule or statistical criterion...". A typical example of aggregate attitude produced by a statistical criterion is



shared belief, namely the fact that all agents (or most of the agents) in a set of agents believe that a certain proposition  $p$  is true. An example of aggregate attitude produced by an aggregation rule is the collective acceptance of a jury about a given proposition  $p$  obtained by majority voting: the jury believes that the proposition  $p$  is true if and only if the majority of the members of the jury has expressed the individual opinion that  $p$  is true. Aggregate attitudes produced by aggregation rules are the objects of analysis of judgement aggregation, an important research area in social sciences and AI (see [54; 83] for an introduction to judgement aggregation). Differently from common attitudes, aggregate attitudes do not require a level of common awareness by the members of the group. That is, a group can hold an aggregate attitude even though the members of the group do not necessarily believe so. For example, the fact that two agents share the belief that  $p$  is true does not necessarily imply that they individually believe that they share this belief. As emphasized by [84] “...a common attitude (of a collective) is an attitude held by all individual members of the collective, where their holding it is a matter of common awareness”, where the term “common awareness” refers to the fact that every member of the group believes that the group has the common attitude, that every member of the group believes that every member of the group believes that the group has the common attitude, and so on. A typical example of common attitude is common belief: every agent in the group believes that  $p$  is true, every agent in the group believes that every agent in the group believes that  $p$  is true, and so on ad infinitum.

**Functions of collective attitudes** Collective attitudes play a crucial role in the society as: (i) they provide the basis of our common understanding through communication, (ii) they ensure coordination between agents, (iii) they are fundamental constituents of collaborative activities between agents acting as members of the same team.

In linguistic, the concept of common ground in a conversation is typically conceived as the common knowledge (or common belief) that the speaker and the hearer have about the rules of the language they use and about the meaning of the expressions uttered by the speaker [136]. Indeed, language use in conversation is a form of social activity that requires a certain level of coordination between what the speaker means and what the addressee understands the speaker to mean. Any utterance of the speaker is in principle ambiguous because the speaker could use it to express a variety of possible meanings. Common ground — as a mass of information and facts mutually believed by the speaker and the addressee — ensures coordination by disambiguating the meaning of the speaker’s utterance. For example, suppose two different operas, “Don Giovanni” by Mozart and “Il Barbiere di Siviglia” by Rossini, are performed in the same evening at two different theaters. Mike goes to see Don Giovanni and the next morning sees Mary and asks “Did you enjoy the opera yesterday?”, identifying the referent of the word “opera” as Don Giovanni. In

order to ensure that Mary will take “opera” as referring to Don Giovanni, it has to be the case that the night before Mary too went to see Don Giovanni, that Mary believes that Mike too went to see Don Giovanni, that Mary believes that Mike believes that Mary too went to see Don Giovanni, and so on.

Moreover, since the seminal work by David Lewis [82], the concept of common belief has been shown to play a central role in the formation and emergence of social conventions.

Finally, collective attitudes such as common goal and joint intention are traditionally used in the philosophical area and in AI to account for the concept of collaborative activity [24; 55; 40; 41]. Notable examples of collaborative activity are the activities of painting a house together, dancing together a tango, or moving a heavy object together. Two or more agents acting together in a collaborative way need to have a common goal and need to form a shared plan aimed at achieving the common goal. In order to make collaboration effective, each agent has to commit to her part in the shared plan and form the corresponding intention to perform her part of the plan. Moreover, she has to monitor the behaviors of the others and, eventually, to reconsider her plan and adapt her behavior to the new circumstances.

**The origin of collective attitudes** Where do collective attitudes come from? How are they formed? There is no single answer to these questions, as collective attitudes can originate in many different ways.

As explained above, aggregate attitudes are the product of aggregation procedures like majority voting or unanimity (cf. [85]). The agents in a certain group decide to use a certain aggregation rule. Then, every agent expresses her opinion about a certain issue  $p$  and the aggregation rule is used to determine what the group believes or what the group accepts. Examples of collective attitudes originating from the aggregation of individual attitudes are group belief and collective acceptance.

Collective attitudes, such as shared belief and common belief, can also be formed through communication or joint perception. A source of information announces to all agents in a group that a certain proposition  $p$  is true. Under the assumption that every agent perceives what the information source says and that every agent in the group trusts the information source’s judgement about  $p$ , the agents will share the belief that  $p$  is true as a result of the announcement. Creation of common belief through communication requires satisfaction of certain conditions that are implicit in the concept of public announcement, as defined in the context of public announcement logic (PAL) [115], the simplest logic in the family of dynamic epistemic logics (DEL) [147]. Specifically, to ensure that an announcement will determine a common belief that the announced fact is true, every agent in the group has to perceive what the information source says, every agent in the group has to perceive that every agent in the group perceives what the information source says, and so on. The latter is called *co-presence* condition in the linguistic literature [30].

The concept of co-presence becomes particularly relevant in the perspective of designing artificial systems situated in a physical environment that need to acquire common belief of certain facts in order to achieve coordination and to make collaboration effective. For example, imagine two robots moving in the physical environment. A source of information signals to them that there is a danger. It does this by emitting a red light. The robots will be able to form different levels of mutual belief about this fact depending on: (i) their spatial positions and the orientation of their sensors with respect to the source of information, and (ii) the perception of the other robots' spatial positions and of the orientations of the other robots' sensors with respect to the source of information. The concept of co-presence applies not only to agents interacting in a physical environment but also to agents interacting in a virtual environment (e.g., virtual characters of a videogame).

**A side note: collective acceptance vs. common belief** A property that clearly distinguishes collective acceptance from common belief is that common belief implies shared belief, while collective acceptance does not: when there is a common belief in a group of agents  $C$  that a certain proposition  $p$  is true then each agent in  $C$  individually believes that  $p$  is true, while it might be the case that there is a collective acceptance in  $C$  that  $p$  is true, and at the same time one or several agents in  $C$  do not individually believe that  $p$  is true. For example, the members of a Parliament might collectively accept (*qua* members of the Parliament) that launching a military action against another country is legitimate because by majority voting the Parliament decided so, even though some of them — who voted against the military intervention — individually believe the contrary. This difference is due to the fact that collective acceptance is a kind of aggregate attitude which can be formed through aggregation procedures others than unanimity.

Another important difference between collective acceptance and common belief is the irreducibility of collective acceptance to the individual level. In particular, it has been emphasized that, while common belief is strongly linked to individual beliefs and can be reduced to them, collective attitudes such as collective acceptance cannot be reduced to a composition of individual attitudes. This aspect is particularly emphasized by Gilbert [47] who follows Durkheim's non-reductionist view of collective attitudes [42]. According to Gilbert, any proper group attitude cannot be defined only as a label on a particular configuration of individual attitudes, as common belief is. In [48; 143] it is suggested that a collective acceptance of a set of agents  $C$  is based on the fact that the agents in  $C$  identify themselves as members of a certain group, institution, team, organization, etc. and recognize each other as members of the same group, institution, team, organization, etc. Common belief and common knowledge, as traditionally defined in epistemic logic [44], do not entail this aspect of mutual recognition and identification with respect to the same group, institution, team, organization, etc.

### 3.2 Grounding institutions and norms on collective attitudes

In the previous section we have explained how collective attitudes are generated from mental attitudes through aggregation procedures, communication or joint perception.

The next step in our analysis is to explain how institutions and norms are grounded on collective attitudes of different types including collective acceptance and common belief. The term “grounded” means that the existence and the evolution of institutions and norms depend on the existence and the evolution of the collective attitudes of the agents who are members of the institution and who are subject to the norm.

We focus here on two forms of grounding that have been considered in the literature: the grounding of institutions on collective acceptance and the grounding of conventions on common belief.

**Collective acceptance and institutions** The problem of understanding what institutions are and how they function has been addressed both in social sciences, in philosophy and in legal theory. Computer scientists working in the area of multi-agent systems have been interested in devising artificial institutions, modeling their dynamics and the different kinds of rules and norms of an institution that agents have to deal with. Following [110, p. 3], artificial institutions can be conceived as “the rules of the game in a society or the humanly devised constraints that structure agents’ interaction”. In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system [43]. More sophisticated models of institutions leave to the agents’ autonomy the decision whether to comply or not with the specified rules and norms of the institution [2; 88]. However, all previous models abstract away from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members.

What these models of artificial institutions neglect is the fundamental relationship between institutions and the collective attitudes of their members and, in particular, the fact that the existence and the dynamics of an institution (norms, rules, institutional facts, etc.) are determined by the collective attitudes of the agents which identify themselves as members of the institution. This aspect is emphasized in the following quote from [101, p. 77]:

“only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions.”  
[Emphasis added].

Prominent philosophical theories of institutional reality conceives collective acceptance as the collective attitude on which institutions are grounded [128; 142]. The relationship

between acceptance and institutions has also been emphasized in the philosophical doctrine of Legal Positivism [62]. According to Hart, the foundations of an institution consist of adherence to, or acceptance of, an ultimate rule of recognition by which the validity of any rule of the institution may be evaluated.<sup>16</sup>

**Common belief and conventions** Convention is a concept that has been widely studied in economics [138], philosophy [16; 141] and computer science [153; 151; 131; 129], given the fundamental role it plays in the regulation of both human and artificial societies.

Eating manners, the kind of clothes we wear in office, and the side of the road on which we drive are mundane examples of convention. Roughly, a social convention is a customary, arbitrary and self-enforcing rule of behavior that is generally followed and expected to be followed in a group or in a society at large [82]. When a social convention is established, everybody behaves in an agreed-upon way even if they did not in fact explicitly agree to behave in this way. A social convention can thus be seen as a kind of tacit agreement that has evolved out of a history of previous interactions [138; 141].

Since the seminal contribution by David Lewis [82], the modern approach to conventions is rooted both in epistemic logic and in evolutionary game theory. The *epistemic approach* to the study of conventions has focused on the characterization of the kind of mutual beliefs and expectations that are required for a group to adopt a certain convention [34; 132; 150] and on the distinction between the epistemic conditions of conventions in contrast with the epistemic conditions of social norms [14]. The epistemic approach clearly highlights the fact that conventions are grounded on collective attitudes. Indeed, according to the well-known definition of convention by David Lewis [82, pp. 76], a given regularity of behavior  $R$  is a convention for a population of agents  $P$  at a recurrent situation  $S$ , only if the agents in the population  $P$  *mutually expect* everyone in  $P$  to conform to the regularity  $R$  in the situation  $S$  (and commonly believe so). In other words, for a convention to exist, the agents in the population have to form a mutual expectation about each other's behavior (and a common belief about this). Consider the example of driving on the left-hand side in the UK. This is a convention as every person in the UK expects other people in the UK to drive on the left-hand side of the road. Moreover, every person in the UK expects other people to drive on the left-hand side of the road *because and as long as* she expects other people to expect everyone to drive on the left-hand side of the road.

The *evolutionary approach* to the study of conventions has focused on the conditions under which a certain convention can emerge on a given population of agents depending on the agents' learning capabilities. Notable examples of this approach are the models by Kandori et al. [71] and Young [156] which make predictions about the conditions under which

---

<sup>16</sup>In Hart's theory, the rule of recognition is the rule that specifies the ultimate criteria of validity in a legal system.

agents converge to equilibrium in a certain coordination game by learning the others' play and adjusting their strategies over time. For instance, Kandori et al.'s model investigates the dynamic process that leads the agents to converge to the risk dominant equilibrium in a repeated  $2 \times 2$  coordination game.

It is worth noting that the epistemic approach and the evolutionary approach to the study of conventions have not yet been reconciled. Indeed, none of the existing evolutionary models of conventions deals with the epistemic aspect of conventions, as they do not assume agents to be cognitive and only consider a simplified notion of convention as a mere regularity of behavior.

### 3.3 Logics for institutions

In [95] a modal logic of collective acceptance is proposed, in accordance with the philosophical theories of this notion discussed in Section 3.2. In the logic, collective acceptance is conceived as the collective attitude that some agents have *qua* members of the same institution. In particular, a collective acceptance held by a set of agents  $C$  *qua* members of a certain institution  $x$  is the kind of acceptance the agents in  $C$  are committed to when they are “functioning together as members of the institution  $x$ ”, that is, when the agents in  $C$  identify and recognize each other as members of the institution  $x$ . For example, in the context of the institution Greenpeace agents (collectively) accept that their mission is to protect the Earth *qua* members of Greenpeace. The state of acceptance *qua* members of Greenpeace is the kind of acceptance these agents are committed to when they are functioning together as members of Greenpeace, that is, when they identify and recognize each other as members of Greenpeace. The logic accounts for different kinds of aggregation procedures that the members of an institution may adopt in order to build a collective acceptance of a given fact. This includes unanimity, majority and a criterion based on leadership according to which what the members of an institution collectively accept coincides with the acceptance of the legislator of the institution. Moreover, the logic clearly distinguishes collective acceptance from common belief, by emphasizing the fact that, while common belief is reducible to individual beliefs, collective acceptance cannot be reduced to individual attitudes of the members of an institution. The fact that collective acceptance is not reducible to individual attitudes is reflected in the formal semantics of the logic. While in epistemic logic common belief is commonly represented by means of the transitive closure of the union of the accessibility relations for the individual beliefs, the accessibility relation for collective acceptance is not definable in terms of the accessibility relations for individual beliefs or individual acceptances. Moreover, collective acceptance entails the notion of “group identification” that is not reducible to the individual level.

Following the idea of some prominent philosophical theories of institutions [128; 142] according to which institutional reality only exists in relation with the collective acceptance

of institutional facts by the members of the institution, a systematic analysis of institutional concepts in the context of this logic is given. This includes the concepts of weak permission, strong permission, obligation and constitutive rule.

The relationship between the logic of collective of acceptance and existing logics of institutions has also been investigated. This includes the comparison between the logic of collective acceptance and the logic of institutional facts proposed by [69] and refined more recently by [53]. According to [69; 53], the primary aspect of institutional facts is their being true in the context of an institution  $x$ .

In [95], the bridge between collective acceptance and informal institutions is built by assuming that:

a certain fact  $\varphi$  is true in the context of an informal institution  $x$  if and only if the members of the informal institution  $x$  collectively accept that  $\varphi$  is true (in the context of  $x$ ).

Differently from formal or legal institutions, informal institutions have no official of the law who is in charge of promulgating new norms and who is the guarantor of their validity. An example of informal institution is a language whose rule specifying the relationship between a certain utterance and its meaning is shared by a group of people: in the context of this group, the utterance has a certain meaning since the language speakers collectively accept this.

In [94], the analysis is extended to formal and legal institutions in which legislators and officials of the law exist who are in charge of either creating new norms or suppressing existing ones out of collective deliberation and who are guarantors of the norms' validity. Specifically, it is assumed that:

a certain fact  $\varphi$  is true in the context of a formal institution  $x$  if and only if the legislators of the institution  $x$  collectively accept that  $\varphi$  is true (in the context of  $x$ ).

For example, according to the French law, the legal drinking age is 18 since this fact is accepted by the French legal authority. As emphasized in Section 3.2, this is close to Hart's idea that a legal norm exists because it adheres to the standards of validity specified by the ultimate rule of recognition that has to be accepted by the legal authority. For example, the Italian legal authority accepts that a norm is valid as far as it has been promulgated by the Italian parliament and published in the "Gazzetta Ufficiale della Repubblica Italiana" (Official Gazette of the Italian Republic).

## 4 Conclusion: closing the circle

In the previous sections we have explained: (i) the role of mental attitudes in decision-making and in action performance as well as the relationship between mental attitudes and emotion (Section 2), (ii) how collective attitudes are generated from mental attitudes as well as the relationship between institutions and norms, on the one hand, and collective attitudes, on the other hand (Section 3). More generally, we have moved from the mental level to the collective level and, then, from the collective level to the institutional-normative level. It is now time to close the circle by going back to mind.

The relevant question here is the following: how do institutions and norms, that are grounded on agents' collective attitudes retroactively influence decision-making and action?

First of all, for a norm or convention to affect an agent's decision, it has to be recognized by the agent, that is, the agent has to believe that the norm or convention exists and that if she does not conform to it, she will incur a violation. The latter is called *normative belief* by [32] (see also [6]). Recognition of a convention is guaranteed, if the agent belongs to the group of agents in which the convention holds. Indeed, as emphasized in Section 3.2, according to Lewis' definition, a certain regularity of behavior  $R$  is a convention for a population of agents  $P$  if and only if the agents in  $P$  mutually expect everyone in  $P$  to conform to the regularity  $R$  and commonly believe so. Thus, if agent  $i$  is a member of  $P$  and  $R$  is convention for  $P$ , then  $i$  has to believe that  $R$  is convention for  $P$ . The latter follows from the fact that if the agents in  $P$  have a common belief that some proposition  $p$  holds, then every agent in  $P$  has to believe so.<sup>17</sup>

Once the norm or convention with its associated costs and sanction for violation has been recognized by an agent, the agent will take it into consideration in her decision-making process. For the sake of clarity, we here distinguish *norm compliance* from mere *norm following*. Norm compliance requires the *goal* to conform to the content of the norm. In other words, for an agent to comply with a norm, she has to be motivated by the goal of conforming to what the norm prescribes. For example, an agent complies with the norm of paying taxes if she wants to pay taxes, after having recognized the corresponding norm that she ought to pay taxes. Norm following just requires that the agent chooses an action *knowing that* this choice will lead her to conform to what the norm prescribes. To sum up, while norm compliance requires *purposively* (or *intentionally*) conforming to what the norm prescribes, norm following only requires *knowingly* conforming to what the norm prescribes. Under the assumption that "purposively doing" implies "knowingly doing", norm compliance can be seen as a special case of norm following.

Two different forms of norm compliance exist. As we have emphasized in Section 2.1, some norms are internalized by the agent and give rise to moral values. If the agent decides

---

<sup>17</sup>This property can be formally proved in the logic of common belief [44].



to comply with them, she does it for ethical or moral reasons. In these cases, the agent's goal of conforming to what the norm prescribes is mainly originated from moral considerations. This is *ethical or moral compliance*. For example, an agent may comply with the legal obligation to pay taxes for ethical or moral reasons: the agent wants to pay taxes because she is motivated by the moral value to behave honestly. More generally, ethical compliance requires that the agent's goal of conforming to what the norm prescribes does not depend on the agent's actual desires<sup>18</sup> but only on the agents' actual moral values.<sup>19</sup>

In other cases, the agent complies with the norm because she desires to avoid the sanction or the social cost as a consequence of the violation and because she fears punishment. This is *opportunistic compliance* which is typical for conventions such as the following one:

Except for pizza, sandwiches and other "finger foods", don't eat with your fingers.

This is a convention in Europe, as every person in Europe expects other people in Europe to follow it and every group of European people has a common belief that each of them expects the others to follow the convention. An European person believes that the convention exists and wants to follow it because she desires to avoid the social cost associated with the violation (e.g., the cost of being publicly blamed if she eats the food with her fingers).

In the case of opportunistic compliance, the agent wants to conform with what the norm prescribes because the consequences of norm violation (e.g., sanction, social cost, punishment) are undesirable for her, while the consequences of norm fulfillment (e.g., reward, social approval) are desirable for her. More generally, opportunistic compliance requires that the agent's goal of conforming to what the norm prescribes does not depend on the agent's actual moral values but only on the agents' actual desires.

We conclude the paper with the general observation that, although norm compliance has been extensively studied in the area of multi-agent systems, with an emphasis on both its logical aspects [3; 122; 74], and computational aspects [33; 4; 88; 149; 79], there is still no formal model which captures the distinctions between norm following and norm compliance, and between ethical compliance and opportunistic compliance. We believe this is an important issue. Its understanding would allow to complement a bottom-up approach to institutions, grounding them on the mental level via the collective level, with a top-down approach, explaining how institutions and norms influence the agents' cognition.

---

<sup>18</sup>This means that if the agent did have different desires in her mind, he would have had still the goal to follow the norm.

<sup>19</sup>This means that it is possible for the agent to reconsider her actual moral values in such a way that her goal to follow the norm is also reconsidered.

## References

- [1] C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, 2009.
- [2] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Quantified coalition logic. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1181–1186. AAAI Press, 2007.
- [3] T. Ågotnes, W. van der Hoek, and M. Wooldridge. Robust normative systems and a logic of norm compliance. *Logic Journal of the IGPL*, 18(1):4–30, 2009.
- [4] N. Alechina, M. Dastani, and B. Logan. Programming norm-aware agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, pages 1057–1064. ACM Press, 2012.
- [5] R. Alur, T. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49:672–713, 2002.
- [6] G. Andrighetto, M. Campennì, F. Cecconi, and R. Conte. The complex loop of norm emergence: a simulation model. In K. Takadama, C. C. Revilla, and G. Deffuant, editors, *The Second World Congress on Social Simulation*, LNAI. Springer-Verlag, 2010.
- [7] G. E. M. Anscombe. *Intention*. Basil Blackwell, 1957.
- [8] R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.
- [9] A. Baltag, S. Smets, and J. A. Zvesper. Keep aðhoppingar for rationality: a solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.
- [10] P. Battigalli and M. Dufwenberg. Guilt in games. *The American Economic Review*, 97(2):170–176, 2007.
- [11] P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *J. of Economic Theory*, 106(2):356–391, 2002.
- [12] R. F. Baumeister, A. M. Stillwell, and T. F. Heatherton. Guilt: an interpersonal approach. *Psychological Bulletin*, 115(2):243–267, 1994.
- [13] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.
- [14] C. Bicchieri. *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press, 2006.
- [15] K. Binmore. *Fun and Games: A Text on Game Theory*. D. C. Heath and Company, 1991.
- [16] K. Binmore. *Natural Justice*. Oxford University Press, 2005.
- [17] A. Bjorndahl, J. Y. Halpern, and R. Pass. Axiomatizing rationality. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning: (KR 2014)*. AAAI Press, 2014.
- [18] C. Boutilier. Towards a logic for qualitative decision theory. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 75–86. AAAI Press, 1994.
- [19] R. I. Brafman and Moshe Tennenholtz. An axiomatic treatment of three qualitative decision

- criteria. *Journal of the ACM*, 47(3):452–482.
- [20] R. I. Brafman and Moshe Tennenholtz. On the foundations of qualitative decision theory. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'96)*, pages 1291–1296. AAAI Press, 1996.
- [21] A. Brandenburger, A. Friedenberg, and J. Keisler. Admissibility in games. *Econometrica*, 76:307–352, 2008.
- [22] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.
- [23] M. Bratman, D. J. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.
- [24] M. Bratman. Shared cooperative activity. *The Philosophical Review*, 101(2):327–41, 1992.
- [25] J. Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011.
- [26] J. Broersen, M. Dastani, J. Hulstijn, and L. van der Torre. Goal generation in the boid architecture. *Cognitive Science Quarterly*, 2(3-4):428–447, 2002.
- [27] A. Caplin and J Leahy. Psychological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, 116(1):55–79, 2001.
- [28] Cristiano Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103:157–182, 1998.
- [29] G. Charness and M. Dufwenberg. Guilt in games. *Econometrica*, 74(6):1579–1601, 2009.
- [30] H. Clark and C. Marshall. Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, and I. A. Sag, editors, *Elements of discourse understanding*. 1981.
- [31] P. R. Cohen and H. J. Levesque. Reasons: Belief support and goal dynamics. *Artificial Intelligence*, 42:213–61, 1990.
- [32] R. Conte and C. Castelfranchi. From conventions to prescriptions. towards an integrated view of norms. *Artificial Intelligence and Law*, 7:323–340, 1999.
- [33] N. Criado Pacheco, E. Argente, P. Noriega, and V. Botti. Human-inspired model for norm compliance decision making. *Information Sciences*, 245:218–239, 2013.
- [34] R. P. Cubitt and R. Sugden. Common knowledge, salience and convention: a reconstruction of david lewis' game theory. *Economics and Philosophy*, 19:175–210, 2003.
- [35] A. Damasio. *Descartes Error: Emotion, Reason and the Human Brain*. Putnam Publishing, New York, 1994.
- [36] M. Dastani and E. Lorini. A logic of emotions: from appraisal to coping.
- [37] D. Davidson. Intending. In *Essays on Actions and Events*. Oxford University Press, New York, 1980.
- [38] D. C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts, 1987.
- [39] F. Dretske. *Explaining behavior: reasons in a world of causes*. MIT Press, 1988.
- [40] B. Dunin-Keplicz and R. Verbrugge. Collective intentions. *Fundamenta Informaticae*, 51(3):271–295, 2002.
- [41] B. Dunin-Keplicz and R. Verbrugge. *Teamwork in Multi-Agent Systems: A Formal Approach*.

Wiley, 2010.

- [42] E. Durkheim. *The rules of Sociological Method*. Free Press, New York, 1982. first published in French in 1895.
- [43] M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, pages 348–366, Berlin, 2001. Springer Verlag.
- [44] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.
- [45] N. H. Frijda, P. Kuipers, and E. Ter Schure. Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, 57(2):212–228, 1989.
- [46] P. Galeazzi and E. Lorini. Epistemic logic meets epistemic game theory: a comparison between multi-agent Kripke models and type spaces. *Synthese*, forthcoming, 2016.
- [47] M. Gilbert. Modelling collective belief. *Synthese*, 73(1):185–204, 1987.
- [48] M. Gilbert. *On Social Facts*. Routledge, London and New York, 1989.
- [49] M. Goldszmidt and J. Pearl. Qualitative probability for default reasoning, belief revision and causal modeling. *Artificial Intelligence*, 84:52–112, 1996.
- [50] R. M. Gordon. *The structure of emotions*. Cambridge University Press, Cambridge, 1987.
- [51] E. Gradel. Model checking games. *Electronic Notes in Theoretical Computer Science*, 67:15–34, 2002.
- [52] J. Gratch and S. Marsella. A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.
- [53] D. Grossi, J.-J. Ch. Meyer, and F. Dignum. Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation*, 16(5):613–643, 2006.
- [54] D. Grossi and G. Pigozzi. *Judgment Aggregation: A Primer*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2014.
- [55] Barbara Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.
- [56] J. Haidt. The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of affective sciences*, pages 852–870. 2003.
- [57] A. Hájek and P. Pettit. A theory of human motivation. *Australian Journal of Philosophy*, 82:77–92, 2004.
- [58] J. Y. Halpern. Beyond nash equilibrium: Solution concepts for the 21st century. In K. R. Apt and E. Gradel, editors, *Lectures in Game Theory for Computer Scientists*, pages 264–289. 2011.
- [59] J. Y. Halpern and R. Pass. A logical characterization of iterated admissibility. In A. Heifetz, editor, *Proc. of TARK 2009*, pages 146–155, 2009.
- [60] J. Harsanyi. Morality and the theory of rational behaviour. In A.K. Sen and B. Williams, editors, *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, 1982.
- [61] J. C. Harsanyi. Games with incomplete information played by ‘bayesian’ players. *Management Science*, 14:159–182, 1967.

- [62] H. L. A. Hart. *The concept of law*. Clarendon Press, Oxford, 1992. new edition.
- [63] A. Herzig and D. Longin. C&L intention revisited. In *Proceedings of the 9th International Conference on Principles on Principles of Knowledge Representation and Reasoning (KR 2004)*, pages 527–535. AAAI Press, 2004.
- [64] J. Hintikka and G. Sandu. Game-theoretical semantics. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 361–410. Elsevier, 1997.
- [65] A. Hopfensitz and E. Reuben. The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, 119(540):1534–1559, 2009.
- [66] J. F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.
- [67] I. L. Humberstone. Direction of fit. *Mind*, 101(401):59–83, 1992.
- [68] J. Doyle and R. Thomason. Background to qualitative decision theory. *The AI Magazine*, 20(2):55–68, 1999.
- [69] A. Jones and M. J. Sergot. A formal characterization institutionalised power. *Journal of the IGPL*, 4:429–445, 1996.
- [70] D. Kahneman and D. T. Miller. Norm theory: comparing reality to its alternatives. *Psychological Review*, 93(2):136–153, 1986.
- [71] M. Kandori, G. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61:29–56, 1993.
- [72] L. Keiff. Dialogical logic. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2011.
- [73] D. Klein and E. Pacuit. Changing types: Information dynamics for qualitative type spaces. *Studia Logica*, 102:297–319, 2014.
- [74] M. Knobbout and M. Dastani. Reasoning under compliance assumptions in normative multi-agent systems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, pages 331–340. ACM Press, 2012.
- [75] K. Konolige and M. E. Pollack. A representationalist theory of intention. In R. Bajcsy, editor, *Proceedings 13th International Joint Conference on Artificial Intelligence (IJCAI 93)*, pages 390–395, San Francisco, CA, 1993. Morgan Kaufmann Publishers.
- [76] S. Kraus and D. J. Lehmann. Knowledge, belief and time. *Theoretical Computer Science*, 58:155–174, 1988.
- [77] R. S. Lazarus. *Emotion and adaptation*. Oxford University Press, New York, 1991.
- [78] J. LeDoux. *The emotional Brain*. Simon and Schuster, New York, 1996.
- [79] J. Lee, J. Padget, B. Logan, D. Dybalova, and N. Alechina. Run-time norm compliance in BDI agents. In *Proceedings of the Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems (AAMAS 2014)*, pages 1581–1582. ACM Press, 2014.
- [80] D. Lewis. Desire as belief. *Mind*, 97:323–332, 1988.
- [81] D. Lewis. Desire as belief ii. *Mind*, 105:303–313, 1996.
- [82] D. K. Lewis. *Convention: a philosophical study*. Harvard University Press, Cambridge, 1969.
- [83] C. List. The theory of judgment aggregation: an introductory review. *Synthese*, 187(1):179–

207, 2012.

- [84] C. List. Three kinds of collective attitudes. *Erkenntnis*, 79(9):1601–1622, 2014.
- [85] C. List and P. Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.
- [86] J. Locke. *An essay concerning human understanding*. Clarendon Press, Oxford, 1989.
- [87] G. Loomes and R. Sugden. Testing for regret and disappointment in choice under uncertainty. *Economic J.*, 97:118–129, 1987.
- [88] F. Lopez y Lopez, M. Luck, and M. d’Inverno. Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS’04)*, pages 732–739. ACM Press, 2004.
- [89] E. Lorini. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4):372–399, 2013.
- [90] E. Lorini. A logic for reasoning about moral agents. *Logique et Analyse*, 58(230):177–218, 2016.
- [91] E. Lorini. A minimal logic for interactive epistemology. *Synthese*, 193(3):725–755, 2016.
- [92] E. Lorini and C. Castelfranchi. The cognitive structure of Surprise: looking for basic principles. *Topoi: An International Review of Philosophy*, 26((1)):133–149, 2007.
- [93] E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45–77, 2008.
- [94] E. Lorini and D. Longin. A logical account of institutions: from acceptances to norms via legislators. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 38–48. AAAI Press, 2008.
- [95] E. Lorini, D. Longin, B. Gaudou, and A. Herzig. The logic of acceptance: grounding institutions on agents’ attitudes. *Journal of Logic and Computation*, 19(6):901–940, 2009.
- [96] E. Lorini, D. Longin, and E. Mayor. A logical analysis of responsibility attribution : emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6):1313–1339, 2014.
- [97] E. Lorini and R. Muehlenbernd. The long-term benefits of following fairness norms: a game-theoretic analysis. In *Proceedings of the 18th Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2015)*, pages 301–318, Berlin, 2015. Springer-Verlag.
- [98] E. Lorini and F. Schwarzentruber. A Modal Logic of Epistemic Games. *Games, Epistemic Game Theory and Modal Logic*, 1(4):478–526, 2010.
- [99] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4):814–847, 2011.
- [100] K. Ludwig and M. Jankovic. Collective intentionality. In L. McIntyre and A. Rosenberg, editors, *The Routledge Companion to the Philosophy of Social Science*. Routledge, New York, 2016.
- [101] C. Mantzavinos, D.C. North, and S. Shariq. Learning, institutions, and economic performance. *Perspectives on Politics*, 2:75–84, 2004.
- [102] J. G. March and H. A. Simon. *Organizations*. Wiley, New York, 1958.
- [103] A. H. Maslow. A theory of human motivation. *Psychological Review*, 50:370–396, 1943.
- [104] H. McCann. Settled objectives and rational constraints. *American Philosophical Quarterly*,

28:25–36, 1991.

- [105] A. R. Mele. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, Oxford, 1992.
- [106] J.-J. Ch. Meyer. Reasoning about emotional agents. *International J. of Intelligent Systems*, 21(6):601–619, 2006.
- [107] J. J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
- [108] K. Miller and G. Sandu. Weak commitments. In G. Holmström-Hintikka and R. Tuomela, editors, *Contemporary Action Theory, vol.2: Social Action*. Kluwer Academic Publishers, Dordrecht, 1997.
- [109] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.
- [110] D.C. North. *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge, 1990.
- [111] A. Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.
- [112] M. Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.
- [113] A. Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, 2012.
- [114] M. Platts. *Ways of meaning*. Routledge and Kegan Paul, 1979.
- [115] J. A. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 201-216, 1989.
- [116] Z. Pylyshyn. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, Cambridge, Massachusetts, 1984.
- [117] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI architecture. In *Proceedings of KR'91*, San Francisco, CA, 1991. Morgan Kaufmann Publishers.
- [118] R. Reisenzein. Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, 1(3):214–222, 2009.
- [119] N. Rescher. Semantic foundations for the logic of preference. In N. Rescher, editor, *The logic of decision and action*. University of Pittsburgh Press, 1967.
- [120] S. Rick and G. Loewenstein. The role of emotion in economic behavior. In M. Lewis, J. Haviland-Jones, and L. Feldman-Barrett, editors, *The Handbook of Emotion*. Guilford, New York, 2008.
- [121] I.J. Roseman, A.A. Antoniou, and P.E. Jose. Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10:241–277, 1996.
- [122] A. Rotolo. Norm compliance of rule-based cognitive agents. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, pages 2716–2721. AAAI Press, 2011.
- [123] K. R. Scherer, A. Schorr, and T. Johnstone, editors. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, Oxford, 2001.

- [124] T. Schroeder. *Three faces of desires*. Oxford University Press, 2004.
- [125] J. Searle. *Expression and meaning*. Cambridge University Press, 1979.
- [126] J. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, 1983.
- [127] J. Searle. *Rationality in Action*. MIT Press, Cambridge, 2001.
- [128] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.
- [129] S. Sen and S. Airiau. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1507–1512. ACM Press, 2007.
- [130] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.
- [131] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1-2):139–166, 1997.
- [132] G. Sillari. A logical framework for convention. *Synthese*, 147(2):379–400, 2005.
- [133] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956.
- [134] M. Singh and N. Asher. A logic of intentions and beliefs. *Journal of Philosophical Logic*, 22:513–544, 1993.
- [135] R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.
- [136] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25(5-6):701–721, 2002.
- [137] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.
- [138] R. Sugden. *Economics of rights, co-operation and welfare (2nd Edition)*. Palgrave Macmillan, 2004.
- [139] J. P. Tangney. Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, 38(8):1132–1145, 1995.
- [140] D. P. Tollefsen. Collective intentionality and the social sciences. *Philosophy of the Social Sciences*, 32(1):25–50, 2002.
- [141] L. Tummolini, G. Andrighetto, C. Castelfranchi, and R. Conte. A convention or (tacit) agreement betwixt us: on reliance and its normative consequences. *Synthese*, 190(4):585–618, 2013.
- [142] R. Tuomela. *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge, 2002.
- [143] R. Tuomela. *The Philosophy of Sociality*. Oxford University Press, Oxford, 2007.
- [144] P. Turrini, J.-J. Ch. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a Dynamic Logic Account. *Journal Autonomous Agents and Multi-Agent Systems*, 20(3):401–420, 2010.
- [145] J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.
- [146] J. van Benthem, P. Girard, and O. Roy. Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of Philosophical Logic*, 38:83–125, 2009.



- [147] H. P. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Kluwer Academic Publishers, 2007.
- [148] B. van Linder, van der Hoek, and J.-J. Ch. W., Meyer. Formalising abilities and opportunities. *Fundamenta Informaticae*, 34:53–101, 1998.
- [149] M. B. van Riemsdijk, L. A. Dennis, M. Fisher, and K. V. Hindriks. Agent reasoning for norm compliance: a semantic approach. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems (AAMAS 2013)*, pages 499–506. ACM Press, 2013.
- [150] P. Vanderschraaf. Convention as correlated equilibrium. *Erkenntnis*, 42(1):65–87, 1995.
- [151] D. Villatoro, S. Sen, and J. Sabater-Mir. Exploring the dimensions of convention emergence in multiagent systems. *Advances in Complex Systems*, 14(2):201–227, 2011.
- [152] G. H. Von Wright. *The logic of preference*. Edinburgh University Press, 1963.
- [153] W. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 384–389. AAAI Press, 1995.
- [154] M. P. Wellman and J. Doyle. Preferential semantics for goals. In *Proceedings of the Ninth National conference on Artificial intelligence (AAAI'91)*, pages 698–703. AAAI Press, 2001.
- [155] M. Wooldridge. *Reasoning about Rational Agents*. MIT Press, Cambridge, 2000.
- [156] H. P. Young. The evolution of conventions. *Econometrica*, 61:57–84, 1993.
- [157] M. Zeelenberg, W. van Dijk, A. S. R. Manstead, and J. van der Pligt. On bad decisions and disconfirmed expectancies: the psychology of regret and disappointment. *Cognition and Emotion*, 14(4):521–541, 2000.
- [158] J. A. Zvesper. *Playing with Information*. PhD thesis, University of Amsterdam, The Netherlands, 2010.