Synthese (2009) 168:201–248
Knowledge, Rationality & Action 511–558
DOI 10.1007/s11229-009-9460-9

A logical formalization of the OCC theory of emotions

C. Adam · A. Herzig · D. Longin

Received: 22 March 2007 / Accepted: 14 January 2009 / Published online: 24 February 2009

Abstract In this paper, we provide a logical formalization of the emotion triggering process and of its relationship with mental attitudes, as described in Ortony, Clore, and Collins's theory. We argue that modal logics are particularly adapted to represent agents' mental attitudes and to reason about them, and use a specific modal logic that we call Logic of Emotions in order to provide logical definitions of all but two of their 22 emotions. While these definitions may be subject to debate, we show that they allow to reason about emotions and to draw interesting conclusions from the theory.

Keywords Modal logics · BDI agents · Emotions · OCC theory

1 Introduction

There is a great amount of work concerning emotions in various disciplines such as philosophy (Gordon 1987; Solomon and Calhoun 1984), economy (Elster 1998; Loewenstein 2000), neuroscience and psychology. In neuroscience, experiments have highlighted that individuals who do not feel emotions e.g. due to brain damage are unable to make rational decisions (see Damasio 1994 for instance), refuting the

C. Adam (✉)
RMIT University, Melbourne, VIC, Australia
e-mail: carole.adam.rmit@gmail.com

A. Herzig · D. Longin
Institut de Recherche en Informatique de Toulouse, Université de Toulouse, CNRS, Toulouse, France
e-mail: andreas.herzig@irit.fr

D. Longin
e-mail: dominique.longin@irit.fr

commonsensical assumption that emotions prevent agents from being rational. Psychology provides elaborated theories of emotions ranging from their classification (Ekman 1992; Darwin 1872) to their triggering conditions (Lazarus 1991; Ortony et al. 1988) and their impact on various cognitive processes (Forgas 1995).

Computer scientists investigate the expression and recognition of emotion in order to design anthropomorphic systems that can interact with human users in a multi-modal way. Such systems are justified by the various forms of 'anthropomorphic behavior' that users ascribe to artifacts. This has lead to an increasing interest in Affective Computing, with particular focus on embodied agents (de Rosis et al. 2003), ambient intelligence (Bartneck 2002), intelligent agents (Steunebrink et al. 2007), etc. All these approaches generally aim at giving computers extended capacities for enhanced functionality or more credibility. Intelligent embodied conversational agents (ECAs) use a model of emotions both to simulate the user's emotion and to show their affective state and personality. Bates has argued for the importance of emotions to make artificial agents more *believable*: "*It does not mean an honest or reliable character, but one that provides the illusion of life, and thus permits the audience's suspension of disbelief.*" (Bates 1994, p. 122). Indeed, there are many pieces of evidence suggesting that virtual agents and robots (interacting with humans) that are capable to display emotions, to recognize the human users' emotions, and to respond to their emotions in an appropriate way, allow to induce positive feelings in the humans during the interaction and to improve their performance. For instance it has been shown that emotions affect learning (Bower 1992), so many computer scientists have added human-provided emotional scaffolding to their computer tutoring systems in order to increase both student persistence and commitment (Aist et al. 2002) and to improve learning (Elliott et al. 1999). In the same way, other researches show that machines which express emotions and provide emotional feedback to the user, allow to enhance the user's enjoyment (Bartneck 2002; Prendinger and Ishizuka 2005), her engagement (Klein et al. 1999) and performance in task achievement (Partala and Surakka 2004), her perception of the machine (Brave et al. 2005; Picard and Liu 2007) and can engage in more natural dialogs with her (Becker et al. 2004).

The great majority of these works are founded on psychological works about emotion. "What is the best theory of emotion today?" is a question where currently there is no consensus. A theory widely used by computer scientists is the one proposed by Ortony, Clore, and Collins (OCC henceforth). A reason is that this theory is relatively understandable by computer scientists, because it is founded on a combinatory approach of a finite set of criteria allowing to characterize emotions. (We are going to present this theory in more detail in Sect. 2.2, and are going to present more arguments in Sect. 2.4.)

OCC theory provides what may be called a semi-formal description language of emotion types. It neither accounts for relationships between the different components of emotions nor relationships between agents' emotions and their actions. The aim of this paper is to fill this gap by formalizing OCC theory with the help of a language describing agents' mental attitudes such as beliefs, goals or desires. In this way we stay as close as possible to the original psychological theory. More precisely, we aim at modelling the triggering process of emotions in intelligent agents endowed with mental states (i.e. a set of mental attitudes about some contents). What we do is to

describe how a given mental state contributes to the triggering of a given emotion. This problem has to be solved before formalizing the subsequent influence of emotions on any mental process and in particular on planning. In this paper we therefore focus on the influence of mental states on emotions, and do not address the influence of emotions on mental states.

Our aim is to model emotion in a logic of mental attitudes. Formal logic provides a universal vocabulary with a clear semantics and it allows reasoning, planning and explanation of an agent's behavior. A given formal definition of emotions may be criticized, but it still has the advantage to be unambiguous and to allow analysis and verification. In particular, all logical consequences of formal principles must remain intuitive: a logical formalization may reveal consequences (and even inconsistencies) that were 'hidden' in the theory and did not appear before. Formal definitions clearly articulate assumptions and allow to formally derive consequences of certain assumptions: they allow to clearly and concisely articulate the assumptions of a theory and to readily uncover the consequences. All in all, logical formalization is a well-defined scientific program to move forward and develop more widely accepted and clearly defined models.

The logic used here is a particular modal logic that grounds on the philosophy of language, of mind, and of action (Bratman 1987; Searle 1969, 1983), and proposes to model agents via some key concepts such as *mental attitudes* (belief, preference, desirability), action and time. This framework is very close to those commonly used in the agent community and offers well-known interesting features: great explanatory power, formal verifiability, and a rigorous and well-established theoretical frame (from the point of view of both philosophy and formal logic). Note that we are not concerned at this stage with optimizations of our logical theory in view of particular applications; for the time being we leave this to agent designers who might use our model as a basis for their work.

Our aim is also to model emotion in a way that is as faithful as possible to psychology. Thus we believe that our logical theory is built on solid grounds given that OCC theory is a well-established psychological theory. The properties of our logic may be evaluated with respect to the following criteria: (1) the number and types of the emotions that are covered; (2) the examples given by psychologists (is our formalism able to account for these examples?); (3) the theorems following from our model (are these theorems intuitive and relevant? are they in accordance with the formalized psychological theory or do they run counter to it? etc.).

We also believe that the other way round our logic, thanks to its faithfulness to the OCC theory, may contribute to the assessment of this theory. For example the consistency of our logic demonstrates that OCC theory is free of contradictions.

In the rest of the paper, we expose the OCC theory underlying our work (Sect. 2). In Sect. 3 we introduce our logical framework. In Sects. 4 and 5 we detail the event-based and agent-based branches of the OCC theory and their formalization. In Sect. 6 we expose some theorems concerning emotions, particularly relating to causal and temporal links between them. In order not to overload the paper, the proofs of these theorems are gathered in the appendix. In Sect. 7 we discuss some existing logical models of emotions.

2 Emotion theories

To ensure the accuracy of a computational model of emotions, it is important to start from acknowledged psychological theories. There exist several kinds of psychological models of emotions: *evolutionist models* (e.g. Darwin 1872) that are mainly descriptive, giving taxonomies of basic emotions; *dimensional models* (e.g. Russell 1997) that assume that all emotions are similar phenomena, only varying on the values of some dimensions like valence or arousal; these models were sometimes used to describe the dynamics of the expression of emotions (e.g. Becker et al. 2004); *cognitive appraisal theories* (e.g. Ortony et al. 1988) that focus on the cognitive determination of emotions and on their adaptive function.

The concept of appraisal was first introduced by Arnold (1960) to describe the triggering of emotions, together with the concept of action tendencies describing their effects. These two concepts were then studied in many approaches; we present some of the most important ones here (Sect. 2.3), in particular that of OCC theory (Sect. 2.2). Before that, let us first shortly speak about relationships between emotion and cognition through the concept of Intentionality (Sect. 2.1).

2.1 Emotion, cognition, intentionality, and logic

The use of logic for emotion formalization may appear surprising at first glance, and one might consider that they cannot be married. Nevertheless, today the great majority of psychologists work with approaches where emotion and cognition are strongly connected (see Lazarus 1999 for instance: for him, emotion is a part of cognition). Logic can deal with cognition through the well-known BDI logics (for belief, desires, intentions). Cognition refers, among other things, to mental states and reasoning about them. Thus, to say that emotion is in cognition means that emotion is concerned with mental states.

In our view, emotions are always about a state of affairs of the world. In other words, emotion is an *Intentional* components of our mind in the sense of Searle (1983, p. 1): "*Intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world*". (Note that *intention* is just a particular form of Intentionality. To avoid confusions and following Searle (1983), we write "Intentionality" with an upper case letter.) For instance, belief and preference are among the Intentional mental states. Note that only some, but not all, mental states have Intentionality: for instance, forms of nervousness, elation, and undirected anxiety, that are diffuse without any clear link with an object of the world, are not Intentional states. Searle (1983, pp. 29–36) has described some emotions as complex mental attitudes that can be expressed as a combination of beliefs and desires. We want to generalize this approach by applying it to OCC theory. In this perspective, the description of an emotion as a combination of beliefs and desires presupposes that the emotion under concern is an Intentional mental attitude. We therefore do not deal in this paper with other emotional states that are closer to *mood* or that are not *Intentional emotions*, in the sense that they are not concerned with or based on Intentional mental attitudes.

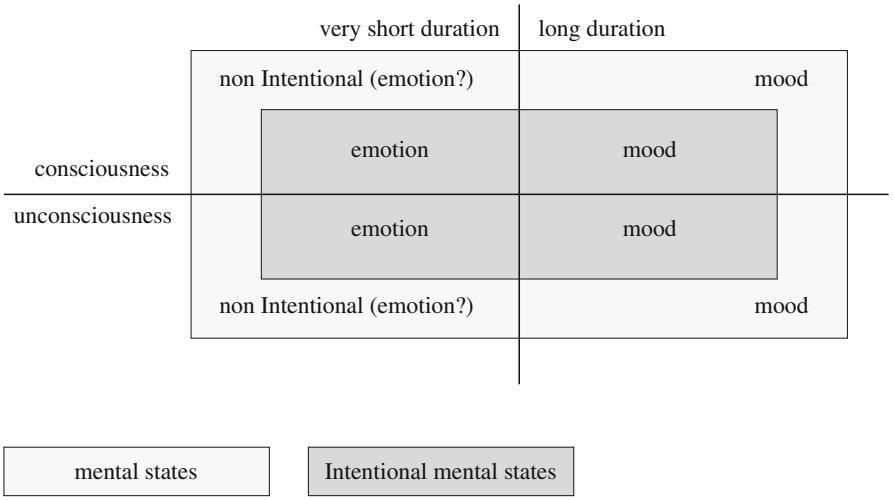

Fig. 1 Emotions, mood, Intentional states, and duration

Following the great majority of psychologists, another difference between emotion and mood is that emotion has a very short duration in time. (See for instance Ortony et al. 1988; Lazarus 1991.) Thus, we can expect that an affective state having a long duration is not so much emotion as mood.

Finally, Intentional mental attitudes can be either conscious or unconscious, and this property is not related to Intentionality. (Searle shows that one can be conscious of non Intentional mental states, and conversely one can be unconscious of an Intentional state, see Searle (1983, Chap. 1).) Figure 1 pictures the situation.

Ortony and colleagues agree that "individual emotions can be specified in terms of personal or interpersonal situation descriptions that are sufficient to produce them" (Ortony et al. 1988, p. 3). More precisely they assume that "if the described situation contains the eliciting condition for a particular emotion, the experience of that emotion can be inferred" (Ortony et al. 1988, p. 3). This is clearly a cognitive approach where emotions are Intentional concepts. This supports our fundamental design choice: emotions can only occur in particular mental states formalized through the logical definition of the mental attitudes constituting their elicitation conditions. Our choice of a logic of mental attitudes is therefore justified both by the fact that it is an appropriate formalization of mental states (see Cohen and Levesque 1990; Rao and Georgeff 1991, 1992; Sadek 1992; Herzig and Longin 2004, for instance) and by the fact that mental attitudes allow to express emotions.

2.2 Ortony, Clore and Collins's theory of emotion

Ortony et al. (1988) propose a cognitive appraisal theory that is structured as a three-branch typology, corresponding to three kinds of stimuli: consequences of events, actions of agents, and aspects of objects. Each kind of stimulus is appraised w.r.t. one central criterion, called *central appraisal variable*. An individual judges the following:

– the desirability of an event, i.e. the congruence of its consequences with the individual's goals (an event is pleasant if it helps the individual to reach his goal, and unpleasant if it prevents him from reaching his goal);
– the approbation of an action, i.e. its conformity to norms and standards;
– the attraction of an object, i.e. the correspondence of its aspects with the individual's likings.

There are some secondary appraisal variables influencing the intensity of the generated emotion, such as the probability of an event, the degree of responsibility of the author of an action and the amount of effort that was provided.

The OCC typology contains twenty-two emotions types[1] that are grouped in six classes. The first branch contains three classes of emotions triggered by the appraisal of the consequences of an event as to its desirability. *Well-being emotions* (joy, distress) arise when an individual appraises an event that has just occurred while only focusing on the desirability of its consequences for himself. *Fortunes-of-others emotions* (happy for, sorry for, resentment, gloating) arise when an individual appraises an event while focusing on its desirability for another individual. *Prospect-based emotions* such as hope or fear arise when an individual appraises the consequences of a prospected event (namely an event that has not occurred yet but is expected to do so) while focusing on the desirability of its consequences for himself. Other prospect-based emotions such as disappointment, relief, fears-confirmed, and satisfaction arise when an individual appraises an event that has just occurred and that was expected, while focusing on its desirability for himself.

The second branch contains only one class of emotion types (*Attribution emotions*) triggered by the appraisal of an action as to its approval, i.e. its conformity to norms and standards. Thus, pride and shame arise when an individual appraises one of his own actions while focusing only on its approval ('does this action conform to the standards?') and not on its consequences. Admiration and reproach arise when an individual appraises an action of another individual while focusing only on its approval.

An other class, common both to Well-being emotions (first branch of the typology) and Attribution emotions (second branch of the typology) is *Compounds emotions* (*attribution-wellbeing*) (remorse, gratification, gratitude, anger) that arise when an individual appraises an action while focusing both on its approval an on the desirability of its consequences.

Here is a complex example where several of the above emotion types are involved. Suppose you and a friend of yours apply for the same position. You believe your CV is better, but then you learn that your friend got the position because he cheated a bit on his CV (say he over-emphasized his participation in some project and gave to some of his papers a "to appear" status although they are just submitted). According to OCC theory you might then feel (1) disappointed (confirmation-based), (2) happy for your friend (fortune of other), and (3) reproach (attribution emotion). The relative

[1] According to the authors, an emotion type is "a distinct kind of emotion that can be realized in a variety of recognizably related forms" (Ortony et al. 1988, p. 15), for example various intensities or various emphasis. In the sequel of this paper, to simplify the vocabulary we generally use the term "emotion" instead of "emotion type".

importance of these three emotions depends on the secondary appraisal variables, which is something we do not account for in our framework. What we deal with here is whether such emotions can indeed be triggered simultaneously by the same event, i.e. whether the conjunction of these three emotions is consistent in our logic.

Finally, the third branch contains one class of emotions: *attraction emotions* (love, hate), triggered by the appraisal of the aspects of objects w.r.t. the individual's likings.

It is important to notice that the authors of the OCC theory intended it to be used in Artificial Intelligence:

> "(…) we would like to lay the foundation for a computationally tractable model of emotion. In other words, we would like an account of emotion that could in principle be used in an Artificial Intelligence (AI) system that would, for example, be able to reason about emotion."
>
> (Ortony et al. 1988, p. 2)

This aim was pretty much reached since OCC theory is the most popular psychological model of emotions in computer science nowadays, and emotional agents widely employ it (e.g. Elliott (1992), Reilly (1996), de Rosis et al. (2003), Jaques et al. (2004), Ochs et al. (2005)). However, it is not the only one, and we can quote emotional agents based on Frijda's theory (e.g. Staller and Petta (2001)) as well as agents based on Lazarus's theory (e.g. Gratch and Marsella (2004)).

2.3 Other theories

Frijda (1986) focuses on the action tendencies induced by emotions. A stimulus first passes through various steps of evaluation determining its characteristics: causes and consequences, relevance and congruence with interests, coping possibilities, urgency. Depending on the result, a control signal is generated to postpone or interrupt the current action. An action preparation is then created (action plan, action tendency, activation mode) that induces physiological changes, and finally an action is selected and executed. Frijda believes that it is the associated action tendency that differentiates basic emotions from each other. Dastani and Meyer (2006) build on this notion of action tendency to define the effect of four emotions on a rational agent's plans.

Lazarus (1991) presents a relational, motivational, cognitive theory of emotion. According to him, emotions result from the cognitive evaluation (or appraisal) of the interaction between an individual and its environment, w.r.t. his motivations and goals. Lazarus distinguishes between the primary appraisal, assessing the relevance and congruence of the stimulus w.r.t. the individual's well-being (that is, does the stimulus help or threaten one of the individual's goals?), and the secondary appraisal, evaluating the available resources to cope with the stimulus (can the individual do something to remove the threatening stimulus?). These two kinds of appraisal are not sequential: they can be executed in any order. Like Arnold, Lazarus considers that emotions induce action tendencies, that cause physiological modifications in order to help the individual adapting to his/her environment. Lazarus' theory is used in the EMA agent (cf. Gratch and Marsella (2004)) whose acronym is an homage to his book.

Scherer (1987) considers emotions as a multicomponent process, with one cognitive component. He introduces an appraisal process consisting in a sequence of stimulus processing steps, called the *Stimulus Evaluation Checks*. This process sequentially evaluates the novelty and unexpectedness of the stimulus, its intrinsic agreeability, its congruence with the individual's goals, the coping possibilities, and its compatibility with norms. Contrarily to Lazarus (1991), these evaluations are ordered. Later, Scherer (2001) associates to each emotion bodily responses, and in particular facial expressions in terms of *Action Units*. The latter are elements defined by Ekman et al. (2002) to represent the moves of the facial muscles. This theory is thus well-adapted to represent the dynamics of facial expressions of an animated agent (e.g. Grizard and Lisetti (2006)).

2.4 Which theory to choose?

Appraisal theories importantly differ one from another on the appraisal criteria that are used, their order of application, and the precise definitions of emotions based on these criteria. We have chosen OCC theory because the careful study of this theory in comparison with others like Lazarus's indicated that it is better adapted to describe the emotions of a virtual agent for several reasons.

First, OCC theory is widely used in the design of emotional agents because its simplicity and implementability matches computer scientists' expectations and needs: it seems that the combination of OCC's finite set of appraisal variables suffices for current applications.

Second, we completely agree that according to OCC, any emotion must be valenced and this valence must always be the same (Ortony et al. 1988, pp. 29–32). This excludes *de facto* states like surprise (that can be either good or bad, or even neither good nor bad) or "feeling abandoned" (in this state one can be sad, but can also not let this get one down and get one's hope up) from being emotions. (In some works, surprise is considered to be an emotion, see Shaver et al. (1987), Meyer et al. (1997), Lorini and Castelfranchi (2006) or Lorini and Castelfranchi (2007) for examples in recent works, or Ekman's works in 70th and Darwin (1872) for older works.) Besides, the necessity for an emotion to be valenced has also the advantage to provide a clear test to differentiate emotions from close notions that are not valenced. Moreover, valence is something naturally captured by logic, making the OCC theory particularly well adapted for a logical formalization.

Third, OCC theory has a simple and elegant tree structure, and uses concepts that have been well-studied in logic such as beliefs, desires and standards. This makes the formalization task easier. In Searle's view (Searle 1983, Sect. 1.5), every Intentional state is not reducible to belief and desire, but every Intentional state contains a belief, a desire, or both a belief and a desire. So-called BDI logics developed in the field of Artificial Intelligence in the last fifteen years (see Cohen and Levesque (1990); Rao and Georgeff (1991) for instance) offer expressive frameworks to represent agents' mental attitudes such as beliefs and desires (see Meyer et al. (1999); Herzig and Longin (2004) and Lorini and Herzig (2008) for instance) and to reconstruct on their basis the cognitive layer of emotions (see Adam (2007); Adam et al. (2006) and also Steunebrink et al. (2007) for instance).

Finally, OCC theory is quite exhaustive, which is important to design robust and versatile agents, i.e. agents that can emotionally react to a great variety of situations. On the contrary Lazarus' theory is more precise but seems to be less exhaustive (see (Adam, 2007, Chap. 4) for a more detailed comparison). We believe that the logical formalization of both theories will allow to compare them in depth in a close future.

The next section presents the logical framework.

3 Logical framework: the *EL* logic

The formal framework of this article is based on our previous BDI framework (Herzig and Longin 2004) that in turn is based on BDI logics (see for instance Cohen and Levesque (1990); Rao and Georgeff (1991); Sadek (1992)). We minimally extend this standard framework by integrating OCC's appraisal variables. As we have said, we restrict our attention to emotion triggering conditions, disregarding the influence of emotions on beliefs, desires and intentions. OCC's emotion triggering conditions do not refer to the mental attitude of intention, that is therefore not required here. As we are only concerned with event-based emotions and agent-based emotions, we here only need to model the *desirability* and *praiseworthiness* variables of OCC theory. But let us first explain these variables and the choices we made in order to model them.

In OCC theory, desirability is about events and is close to the notion of utility. When an event occurs it can satisfy or interfere with agent's goals, and the desirability *variable* has therefore two *aspects*:

> "one corresponding only to the degree to which the event in question appears to have beneficial (i.e. positively desirable) consequences, and the other corresponding to the degree to which it is perceived as having harmful (i.e. negatively desirable, or undesirable) consequences."
>
> (Ortony et al. 1988, p. 49)

It is thus a valenced variable, and an event can be at the same time desirable and undesirable (with respect to the agent's current goals). Desirability (i.e. the positive *aspect* of the desirability *variable*) only influences positive emotions, whereas undesirability (i.e. the negative *aspect* of the desirability *variable*) only influences negative emotions. It follows that the same event can trigger both positive and negative emotions.

While we agree with OCC theory that the primary event may be both desirable and undesirable,[2] such a feature makes a logical formalization difficult because it requires either a paraconsistent notion of desirability such that "φ is desirable" and "$\neg\varphi$ is desirable" are consistent, or a binary notion of desirability that is relativized

[2] For instance the authors give the example of the death of one's friend suffering from a painful disease; on the one hand the loss of one's friend is undesirable, but on the other hand the end of his suffering is desirable.

to goals. Both options would induce several difficulties, that would distract us from our aims; in particular there is no available logic of the latter binary desirability in the literature. A way out is to shift the focus from desirability of events to desirability of *consequences of events*: when someone says that an event is both desirable and undesirable, we are entitled to ask which aspect of this event is desirable and which is undesirable. When considering consequences of events rather than events themselves we may safely suppose that these consequences are either desirable or undesirable with respect to current goals, but not both. For instance, someone's death can entail both an affective loss (undesirable consequence) and the inheritance of a big amount of money (desirable consequence) Ortony et al. (1988). In this example, clearly, the goal concerned with the desirability of the event (that is, to get a big amount of money, or to be rich) is different from the goal concerned with the undesirability of this event (that is, to keep a loved person alive). Correspondingly, our desirability operators have the following properties:

- desirability and undesirability are about consequences of events (and not about the event itself);
- an event can have several consequences;
- each of these consequences cannot be both desirable and undesirable;
- each of these consequences can be evaluated with respect to goals (that may be either achievement goals or maintenance goals).

We formalize the above example by saying that an emotional loss is undesirable and a big amount of money is desirable, while the friend's death is neither desirable nor undesirable.

Things are similar concerning OCC's *praiseworthiness* variable, which concerns the evaluation of actions performed by agents; this evaluation is with respect to standards and has two aspects: actions can be praiseworthy (when they conform to standards) or blameworthy (when they violate standards). Though, we want to avoid to analyze standards in more depth and do not describe how to construct the two aspects of the praiseworthiness variable. We simply define two types of modal operators: one characterizing the praiseworthiness of consequences of actions, and one characterizing the blameworthiness of consequences of actions. Just as for desirability and undesirability, we will consider that such a consequence cannot be both praiseworthy and blameworthy at the same time.

3.1 Syntax

The syntactic primitives of our logic of emotions *EL* are as follows: a nonempty finite set of agents $AGT = \{i_1, i_2, \ldots\ i_n\}$, a nonempty finite set of atomic events $EVT = \{e_1, e_2, \ldots, e_p\}$, and a nonempty set of atomic propositions $ATM = \{p_1, p_2 \ldots\}$. The variables i, j, $k\ldots$ denote agents; α, β, $\gamma\ldots$ denote events; and p, $q\ldots$ denote propositional letters (atomic propositions). The expression $i{:}e$ represents an event e intentionally caused by agent i. We say that $i{:}e$ is an action that is performed by i.

The language $\mathcal{L}_{EL}$ of the *EL* logic is defined by the following BNF (Backus Naur Form):

$$\varphi ::= \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid Bel_i\varphi \mid Prob_i\varphi \mid Des_i\varphi \\ \mid Idl\ \varphi \mid After_{i:\alpha}\varphi \mid Before_{i:\alpha}\varphi \mid G\varphi \mid H\varphi$$

where p rangers over ATM, i ranges over AGT and $i{:}\alpha$ ranges over $AGT \times EVT$. The classical boolean connectives $\wedge$ (conjunction), $\rightarrow$ (material implication), $\leftrightarrow$ (material equivalence) and $\top$ (tautology) are defined from $\neg$ (negation), $\vee$ (disjunction) and $\bot$ (contradiction) in the usual manner.

$Bel_i\varphi$ reads "agent i believes that φ is true". Belief is understood as subjective knowledge, alias truth in all worlds that are possible for the agent: i does not doubt. For instance, $Bel_{i_1} weatherNice$ represents the fact that, from i_1's point of view the weather is nice: i_1 has no doubt about the truth of this fact (but may be wrong).

$Prob_i\varphi$ reads "agent i believes that φ is more probable than $\neg\varphi$", or "i believes that φ probable" for short. This is a weaker form of belief than $Bel_i\varphi$. For example, if agent i_1 is still in bed, $Prob_{i_1} weatherNice$ means that i_1 believes that the weather is probably nice (but i_1 may not be sure about this). What an agent believes is necessarily probable for him, but not the other way round: when i_1 believes that p then p is probable for i_1. (We give more details in the sequel.) Several researchers have investigated logics of probability, mainly in a quantitative (Fagin and Halpern 1994) or comparative way (Segerberg 1971). A few researchers studied a more qualitative notion of probability (Burgess 1969; Herzig 2003), weak belief (Lenzen 1978, 1995) or likelihood (Halpern and Rabin 1987; Halpern and McAllester 1989). All these are based on subjective probability measures. We adopt Burgess's logic, basically because we do not need numbers for our purposes (but they might be added later when investigating particular applications), and because it integrates smoothly with Hintikka's logic of belief that we are going to use.

$Des_i\varphi$ reads "φ is desirable for i". As we have motivated above, instead of desirability of events we here rather deal with desirability of consequences of events. These consequences are evaluated with respect to goals. According to the OCC theory, goals can be either achievement goals (the agent wants to achieve something that is not currently true) or maintenance goals (the agent wants to maintain something that is already true). Moreover, we do not explain the relationship between goals and desirability because goals do not play an explicit role in our definition of emotions. However, goals can be constructed from what is desirable, and intentions can be constructed from goals. (See Herzig and Longin (2004) and Castelfranchi and Paglieri (2007) for more details about such constructions.) In our view, every (achievement or maintenance) goal is about something that is desirable. Thus, if a consequence of an event is (a part of) a goal, then this consequence is desirable. Here, instead of an *occurrent mental attitude* we rather use the notion of *dispositional attitude*, that

corresponds with Bratman's notion of desire and with Cohen & Levesque's notion of goal.[3]

$Idl\ \varphi$ reads "ideally, φ is true". The notion of ideality considered here is taken in a large sense: it embraces all the rules more or less strongly imposed by some authority. They can be strongly explicit (like laws) or more or less implicit (like social or moral rules). When $Idl\ \varphi$ is true then φ is a kind of social preference that is attached to the groups to which the agent belongs. They may therefore differ from the agent's personal preferences. $Idl\ driveRight$, for instance, means that ideally, one drives on the right side of the road, and $Idl\ helpSbInDistress$ means that ideally, one helps somebody in distress.

$After_{i:\alpha}\varphi$ reads "φ will be true after performance of action α by i". This operator allows to describe what is true after the execution of an action, in particular the effects of this action. For instance, $After_{i_1:raiseHand}rightToSpeak_{i_1}$ means that after agent i_1 has raised its hand (say in the classroom) it will have the right to speak. The fact that φ will be true after the performance of action α is conditional on the performance of α: it does not entail that α is currently executed, nor that i intends to execute it. $After_{i:\alpha}\bot$ reads "action α is not executed by agent i". For instance, $After_{i_2:drive}\bot$ means that agent i_2 is not going to drive in the current situation (for instance because i_2 does not have a car).

$Before_{i:\alpha}\varphi$ reads "φ was true before performance of action α by i". It is symmetric to $After_{i:\alpha}$ for the past. $Before_{i:\alpha}\bot$ means "i has not just executed action α". $Before_{i_2:holdsNut}holdsNut_{i_2}$, for instance, means that before crunching a nut, agent i_2 must hold a nut, and $Before_{i_1:drink}\bot$ means that drinking was not i_1's last action.

$G\varphi$ reads "henceforth φ is going to be true". The notion of time that we use here is linear time. It means that states of world are organized in a linear manner, in what is called "histories" in the literature. Thus, $G\varphi$ means that φ is true on the current history from now and everywhere in the future. For instance, $GglassIsBroken$ means that the glass is henceforth broken.

$H\varphi$ reads "φ has always been true in the past". Thus, it means that φ is true on the current history everywhere in the past including now. For instance, $H\neg JohnIsDead$ means that until and including now, John is not dead.

For convenience, we also define the following abbreviations:

$$Happens_{i:\alpha}\varphi \stackrel{def}{=} \neg After_{i:\alpha}\neg\varphi \qquad (\mathrm{Def}_{Happens_{i:\alpha}})$$

$$Done_{i:\alpha}\varphi \stackrel{def}{=} \neg Before_{i:\alpha}\neg\varphi \qquad (\mathrm{Def}_{Done_{i:\alpha}})$$

$$F\varphi \stackrel{def}{=} \neg G\neg\varphi \qquad (\mathrm{Def}_{F})$$

[3] Several concepts of desire exist in the literature. Desire is often viewed as an *occurrent mental attitude*: an attitude that holds here and now, and that is abandoned as soon as it is satisfied, such as an agent's desire on a rainy day that the sun shines, which is dropped when finally the sun comes out. This is similar to Bratman's concept of intention (Bratman 1987) and to Cohen & Levesque's concept of achievement goal (Cohen and Levesque 1990).

$$P\varphi \stackrel{def}{=} \neg H \neg \varphi \qquad (\text{Def}_P)$$

$$Idl_i\varphi \stackrel{def}{=} Bel_i Idl\ \varphi \qquad (\text{Def}_{Idli})$$

$Happens_{i:\alpha}\varphi$ reads "α is about to be performed by agent i, after which φ will be true".[4] In particular, $Happens_{i:\alpha}\top$ reads "action α is about to be performed by agent i". For instance, $Happens_{i_1:tossCoin}(heads \vee tails)$ means i_1 is about to toss a coin, after which the coin will be either heads or tails.

$Done_{i:\alpha}\varphi$ reads "α has just been performed by agent i, and φ was true before" and $Done_{i:\alpha}\top$ reads that agent i has just performed action α. For instance, $Done_{i_2:toDrinkBeer}\, Done_{i_1:toDrinkCoke}\top$ means that agent i_2 has just drunk a beer and just before that, agent i_1 had drunk a coke.

$F\varphi$ reads "φ is true or will be true at some future instant", and $P\varphi$ reads "φ is or was true". For example, $PsunIsShining \wedge FsunIsShining$ means that there is a past instant when the sun was shining and there is a future instant when the sun will be shining.

Finally, $Idl_i\varphi$ reads "from the point of view of the agent i, it is ideal that φ be true". It will be convenient to suppose that it represents an agent's moral norms, that is, the norms that the agent has internalized as true. For instance, $Idli_1$ *be Vegetarian* means that for agent i_1 one should be vegetarian, and $Idli_2\neg(Drunk \wedge Driving)$ means that for agent i_2 it is unideal to drive drunk. Note that in principle not every known ideal (i.e. $Bel_i Idl\ \varphi$) becomes an internalized ideal ($Idl_i\varphi$), i.e., the left to right implication $Bel_i Idl\ \varphi \rightarrow Idl_i\varphi$ of the Definition (Def_{Idl_i}) is not generally valid. The difference is subtle (see Adam et al. (2009) for more details) and here we adopt this simplification because it allows us to avoid an investigation of the relation between internalized and non-internalized standards.

3.2 Semantics

We use a standard Kripke semantics in terms of possible worlds and accessibility relations. The less standard feature is a neighborhood function for the modal operator of probability.

3.2.1 EL frames

At the base of Kripke semantics there is a set of possible worlds W together with accessibility relations for every modal operator. While in most presentations an accessibility relation is a subset of the cartesian product $W \times W$, we here use an equivalent presentation in terms of mappings from W to 2^W.

An EL frame is a 7-tuple $\mathcal{F} = \langle W, \mathscr{B}, \mathscr{P}, \mathscr{D}, \mathscr{I}, \mathscr{A}, \mathscr{G}\rangle$ where:

– W is a set of possible worlds;

[4] Note that the operators *Happens* can be read in this way (which is not their standard dynamic logic reading) because we have supposed determinism: time is linear, entailing that if an action is feasible, then it will happen.

– $\mathscr{B} : AGT \longrightarrow (W \longrightarrow 2^W)$ is the accessibility relation that associates each agent $i \in AGT$ and possible world $w \in W$, with the set $\mathscr{B}_i(w)$ of possible worlds compatible with the beliefs of agent i in w;
– $\mathscr{P} : AGT \longrightarrow (W \longrightarrow 2^{2^W})$ is the function that associates each agent $i \in AGT$ and possible world $w \in W$ with a set of sets of possible worlds $\mathscr{P}_i(w)$ (the *neighborhoods* of w);
– $\mathscr{D} : AGT \longrightarrow (W \longrightarrow 2^W)$ associates each agent $i \in AGT$ and possible world $w \in W$ with the set $\mathscr{D}_i(w)$ of worlds compatible with what is desirable for the agent i in the world w;
– $\mathscr{I} : W \longrightarrow 2^W$ associates each possible world $w \in W$ with the set $\mathscr{I}(w)$ of ideal worlds;
– $\mathscr{A} : AGT \times ACT \longrightarrow (W \longrightarrow 2^W)$ associates each action $i{:}\alpha \in AGT \times ACT$ and possible world $w \in W$ with the set $\mathscr{A}_{i:\alpha}(w)$ of possible worlds resulting from the performance of α by agent i in w;
– $\mathscr{G} : W \longrightarrow 2^W$ associates each possible world $w \in W$ with the set $\mathscr{G}(w)$ of possible worlds in the future of w.

The set $\mathscr{B}_i(w)$ is called a belief state.

3.2.2 Semantical constraints

We impose to our frames the following semantical constraints.

$$\text{All the accessibility relations } \mathscr{B}_i \text{ are serial, transitive and euclidian.} \qquad (SC_1)$$

Thus, belief states are equivalence classes: an agent views several alternative worlds to the real world but cannot distinguish between each of these alternatives. Note that contrarily to knowledge the real world is not necessarily contained in an agent's belief state. Seriality ensures that beliefs are rational: an agent cannot simultaneously believe that p is true and that its negation $\neg p$ is true. Due to the transitivity and euclidianity of the $\mathscr{B}_i$ relations, agents are aware of their beliefs: if $w \in \mathscr{B}_i(w')$ then $\mathscr{B}_i(w) = \mathscr{B}_i(w')$.

If φ is probable for i (i.e. φ is true in all the worlds of some neighborhood, see Chellas (1980, Chap. 7) for more details), then $\neg\varphi$ is not (since each other neighborhood contains at least one world where φ is true). This corresponds to the following constraint:

$$\text{For every } w \in W, \quad \text{if } U_1, U_2 \in \mathscr{P}_i(w) \quad \text{then } U_1 \cap U_2 \neq \emptyset. \qquad (SC_2)$$

Moreover, in order to ensure that at least tautologies are probable, we impose that:

$$\mathscr{P}_i(w) \neq \emptyset \quad \text{for every } w \in W. \qquad (SC_3)$$

Finally, we impose that the neighborhoods in $\mathscr{P}_i(w)$ are subsets of the belief state:

$$\forall U \in \mathscr{P}_i(w), \quad U \neq \emptyset \qquad (SC_4)$$

which entails that belief implies probability.[5]

As explained in the previous section, when desirability is about propositions rather than actions, it is convenient to postulate to consider that desirability is rational: if a proposition is desirable then its converse is not desirable. This is imposed by the following semantical constraint:

$$\text{All the accessibility relations } \mathscr{D}_i \text{ are serial.} \tag{SC$_5$}$$

The situation is similar for ideality: intuitively, the logic of ideality operators is the same as Standard Deontic Logic. (See Åqvist (2002) for more details about Deontic Logic.) Here, the rationality of ideals is justified by the fact that law, moral, habits, standards, etc. are in principle coherent. Thus, if something is ideally true, then its converse must not be true.

$$\text{All the accessibility relations } \mathscr{I} \text{ are serial.} \tag{SC$_6$}$$

Concerning action, we impose that for every $w \in W$:

$$\text{If } w' \in \mathscr{A}_\alpha(w) \quad \text{and} \quad w'' \in \mathscr{A}_\beta(w) \quad \text{then } w' = w''. \tag{SC$_7$}$$

$$\text{If } w \in \mathscr{A}_\alpha(w') \quad \text{and} \quad w \in \mathscr{A}_\beta(w'') \quad \text{then } w' = w''. \tag{SC$_8$}$$

First, this imposes that actions are organized into histories. It does not impede the parallel execution of several actions, but it guarantees that all these parallel actions lead to the same world, i.e., the same time point in the same history. It imposes that all the actions take place in the same history, where the outcome world is the same for all actions performed by all agents.[6] Second, these constraints impose that actions take one time step. Suppose that α and β are performed during the performance of γ. That is: $w' \in \mathscr{A}_\alpha(w)$, $w'' \in \mathscr{A}_\beta(w')$ and $w'' \in \mathscr{A}_\gamma(w)$ hold. Thus, (1) imposes that $w' = w''$, and (2) imposes that $w = w'$, which entails that $w = w' = w''$. Thus, in this case, all actions are reduced to the 'skip' action ('do nothing') and the world remains unchanged. Therefore, actions are deterministic in the future and in the past.

Finally, we impose that the $\mathscr{G}$ accessibility relation is a total preorder,

$$\begin{aligned} &\text{The accessibility relation } \mathscr{G} \text{ and its converse } \mathscr{G}^{-1} \\ &\text{are reflexive, transitive and relatively total:} \\ &\text{if } w_1, w_2 \in \mathscr{G}(w) \quad \text{then } w_1 \in \mathscr{G}(w_2) \text{ or } w_2 \in \mathscr{G}(w_1). \end{aligned} \tag{SC$_9$}$$

[5] Intuitively the elements of $\mathscr{P}_i(w)$ should also be "big" subsets of $\mathscr{B}_i(w)$: every $U \in \mathscr{P}_i(w)$ should contain more elements than its complement $\mathscr{B}_i(w) \backslash U$. But the language of modal logic is not expressive enough to account for this. The above constraint is therefore weaker, and there are neighborhoods satisfying our constraints gathering less than 50% of the worlds (cf. Walley and Fine (1979)). However, this suffices to capture some interesting properties such as inconsistency of some emotions.

[6] This hypothesis does not permit to speak about counterfactual situations such as "if I had done α then φ would be true", but this is not problematic as long as we are not concerned with such hypothetic reasonings.

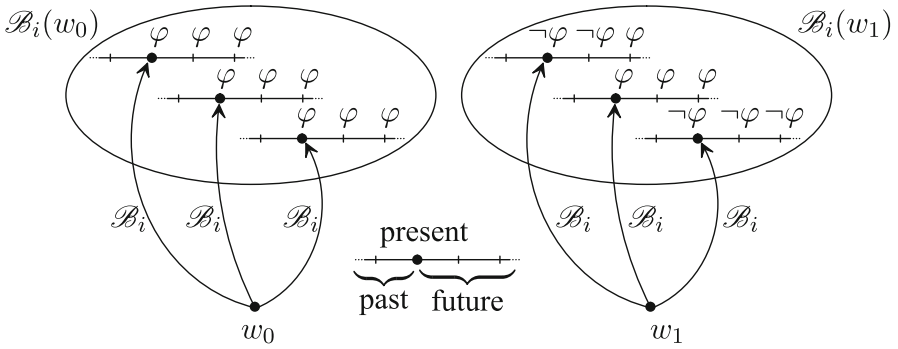

Fig. 2 In world w_0, agent i *believes* that henceforth φ is true; in world w_1, for agent i there are three different possible histories: from top to bottom, the one where φ is currently false but it will occur in the future, the one where φ is henceforth true, and the one where φ is henceforth false

This means that time is linear towards the past (by using $\mathscr{G}^{-1}$) and the future. One might object that at least future should be branching. For us, what is important is not the nature of time in the real world, but rather the perception that agents have of it. Thus, as time is linear here, each world believed to be possible by an agent can be identified with a history, that is a linear sequence of time points, and the diversity of futures is represented through different histories that are possible for the agent at the same world (cf. Fig. 2). In other words, even if time is linear, several futures are possible for the agent, and we therefore have a *subjective version of branching-time*.

Moreover, we impose some constraints involving two or more accessibility relation types. In particular, we suppose that agents are aware of their probabilities and desirabilities, that is, the agents' beliefs about their own subjective probabilities and their desirabilities are correct and complete. We thus impose that, for every $i \in AGT$:

$$\text{if } w' \in \mathscr{B}_i(w) \quad \text{then } \mathscr{P}_i(w) = \mathscr{P}_i(w') \tag{SC$_{10}$}$$

$$\text{if } w' \in \mathscr{B}_i(w) \quad \text{then } \mathscr{D}_i(w) = \mathscr{D}_i(w') \tag{SC$_{11}$}$$

Concerning the relation between belief and action, we suppose that actions are public, in the sense that their occurrence is correctly and completely perceived by all agents. For every $i \in AGT$:

$$\text{if } w' \in \mathscr{B}_i(w) \quad \text{then } (\mathscr{A}_{j:\alpha})^{-1}(w) = \emptyset \quad \text{iff } (\mathscr{A}_{j:\alpha})^{-1}(w') = \emptyset \tag{SC$_{12}$}$$

We also impose that agents do not forget their previous alternatives ("no forgetting", alias "perfect recall" Fagin et al. (1995)). This relies in particular on the preceding hypothesis that actions are public, i.e. that they are perceived correctly and completely by every agent. Thus, for every agent i, $j \in AGT$:

$$\text{if}(\mathscr{B}_i \circ \mathscr{A}_{j:\alpha})(w) \neq \emptyset \quad \text{then } (\mathscr{A}_{j:\alpha} \circ \mathscr{B}_i)(w) \subseteq (\mathscr{B}_i \circ \mathscr{A}_{j:\alpha})(w) \tag{SC$_{13}$}$$

In particular, it is true when i and j are the same agent. In terms of Fig. 2, the agent's belief state after some action was performed at w_0, is a subset of the agent's belief state at w_0 that has been 'progressed' (Reiter 1991) in order to take into account the action occurrence.

Action and time are closely related. In particular, we impose that the future of every world w contains the worlds resulting from the performance of actions in w:

$$\mathscr{G} \supseteq \mathscr{A}_{i:\alpha} \quad \text{for each } i{:}\alpha \in AGT \times EVT \tag{SC$_{14}$}$$

In words, the worlds resulting from the performance of actions in w are necessarily worlds in the future. But the converse is not necessarily true: every world in the future is not necessarily accessible by some action $i{:}\alpha$ in one step: such a hypothesis would be too strong.

For the sake of simplicity, we make the hypothesis that preferences are stable: what is desirable for an agent persists.[7]

$$\text{if } w\mathscr{G}w' \quad \text{then } \mathscr{D}_i(w) = \mathscr{D}_i(w') \tag{SC$_{15}$}$$

This allows to disregard the influence of emotions on desirability. We are aware that our constraint is too strong in the general case, but it is quite realistic for rather short time intervals like a small dialog.

We make the same hypothesis for (social, legal, moral…) obligations, norms, standards… that hold for the agents:

$$\text{if } w\mathscr{G}w' \quad \text{then } \mathscr{I}(w) = \mathscr{I}(w'). \tag{SC$_{16}$}$$

As we do not deal with the dynamics of ideals, it is quite reasonable to consider that ideals are stable, at least for a given time interval.

We call *EL frames* the set of frames satisfying constraints (SC$_1$)–(SC$_{16}$).

3.2.3 EL models and validity

A model $\mathcal{M}$ is a couple $\langle \mathcal{F}, \mathscr{V} \rangle$ where:

- $\mathcal{F}$ is an *EL* frame;
- $\mathscr{V} : W \longrightarrow ATM$ associates each world w with the set $\mathscr{V}_w$ of atomic propositions true in w.

Given a model $\mathcal{M} = \langle \mathcal{F}, \mathscr{V} \rangle$ where $\mathcal{F} = \langle W, \mathscr{B}, \mathscr{P}, \mathscr{D}, \mathscr{I}, \mathscr{A}, \mathscr{G} \rangle$, we recursively define truth of a formula φ at a world w, noted $\mathcal{M}, w \models \varphi$ as follows:

- $\mathcal{M}, w \not\models \bot$;
- $\mathcal{M}, w \models p$ iff $p \in \mathscr{V}_w$;
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$;

[7] This allows concise statements and proofs of theorems (which else would have required the explicit statement of the relevant persistence hypotheses). In recent work we have relaxed this constraint in order to model emotion-focused coping strategies (Adam and Longin 2007).

- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models Bel_i\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{B}_i(w)$;
- $\mathcal{M}, w \models Prob_i\varphi$ iff there exists $U \in P_i(w)$ such that for every $w' \in U, \mathcal{M}, w' \models \varphi$;
- $\mathcal{M}, w \models Des_i\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{D}_i(w)$;
- $\mathcal{M}, w \models Idl\ \varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{I}(w)$;
- $\mathcal{M}, w \models After_{i:\alpha}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{A}_{i:\alpha}(w)$;
- $\mathcal{M}, w \models Before_{i:\alpha}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every w' such that $w \in \mathscr{A}_{ia}(w')$;
- $\mathcal{M}, w \models G\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every $w' \in \mathscr{G}(w)$;
- $\mathcal{M}, w \models H\varphi$ iff $\mathcal{M}, w' \models \varphi$ for every w' such that $w \in \mathscr{G}(w')$.

Validity of a formula φ in the class of all Kripke models obeying our semantic constraints is defined as usual. Thus, φ is true in model $\mathcal{M}$ if and only if $\mathcal{M}, w \models \varphi$ for every w in $\mathcal{M}$. φ is *EL* valid (noted $\models_{EL} \varphi$) if and only if φ is true in every *EL* model $\mathcal{M}$. φ is satisfiable if and only if $\not\models_{EL} \neg\varphi$. φ is a *logical consequence* of a set of (global) hypotheses Γ if and only if for every *EL* model $\mathcal{M}$, if all hypotheses of Γ are true in $\mathcal{M}$ then φ is true in $\mathcal{M}$.

3.3 Axiomatics

We now introduce a set of axioms that our modal operators have to satisfy. All our modal operators except $Prob_i$ are going to be normal modal operators, whose definition we recall first.

3.3.1 Normal operators

$\Box$ is a normal operator iff the axiom (K-$\Box$) and the necessitation rule (RN-$\Box$) hold for $\Box$.

$$\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi) \qquad \text{(K-}\Box\text{)}$$

$$\frac{\varphi}{\Box\varphi} \qquad \text{(RN-}\Box\text{)}$$

In any normal modal logic, the semantics validates the following principles (which are used in some proofs in the appendix):

$$\frac{\varphi \rightarrow \psi}{\Box\varphi \rightarrow \Box\psi} \qquad \text{(RM-}\Box\text{)}$$

$$(\Box\varphi \wedge \Box\psi) \rightarrow \Box(\varphi \wedge \psi) \qquad \text{(C-}\Box\text{)}$$

The dual of $\Box$ is noted $\Diamond$ and obeys the following principle:

$$(\Box\varphi \wedge \Diamond\psi) \rightarrow \Diamond(\varphi \wedge \psi) \qquad \text{(C-}\Box\text{)}$$

and the following inference rule (Chellas 1980, Theorem 4.4, p. 116):

$$\frac{\varphi \rightarrow \psi}{\Diamond\varphi \rightarrow \Diamond\psi} \quad \text{(RM-}\Diamond\text{)}$$

More details on the formal properties of normal modal logics can be found in Chellas (1980, Chap. 4).

3.3.2 Action

$After_\alpha$ and $Before_\alpha$ have the standard tense logic $\mathbf{K}_t$ in a linear time version: a normal modal logic **K** extended with the following axioms (cf. Burgess (2002) for more details):

$$Happens_\alpha\varphi \rightarrow After_\beta\varphi \quad \text{(CD-HA)}$$

$$Done_\alpha\varphi \rightarrow Before_\beta\varphi \quad \text{(CD-DB)}$$

$$\varphi \rightarrow After_\alpha Done_\alpha\varphi \quad \text{(CONV-AD)}$$

$$\varphi \rightarrow Before_\alpha Happens_\alpha\varphi \quad \text{(CONV-BH)}$$

(CD-HA) and (CD-DB) are the axioms of common determinism. For example (CD-HA) means that if an action α is about to happen after which φ, then after any other action β, φ will be true, and similarly in the past for (CD-DB). This entails that actions take one time step, and are deterministic in the future and in the past (one can see that when α is β). The conversion axioms (CONV-AD) and (CONV-BH) link past and future.

Remember that $i{:}\alpha$ reads "agent i does action α".

We highlight here that what we call action is assumed to be intentional, that is the agent always intend to perform actions that he is about to perform. This is the difference between actions and events. Thus if an agents does something unintentionally (like sneezing) it is an event, and it can only trigger event-based emotions. This corresponds to Lazarus' control appraisal variable imposing that one can only reproach something to someone if this person had control over what she did, and to the concept of responsibility in OCC theory (Ortony et al. 1988, p. 54).

3.3.3 Belief

The operators Bel_i have the standard logic **KD45** (cf. Chellas (1980) or Hintikka (1962) for more details). The corresponding axioms are those of normal modal logics plus the following ones:

$$Bel_i\varphi \rightarrow \neg Bel_i\neg\varphi \quad \text{(D-}Bel_i\text{)}$$

$$Bel_i\varphi \rightarrow Bel_i Bel_i\varphi \quad \text{(4-}Bel_i\text{)}$$

$$\neg Bel_i\varphi \rightarrow Bel_i\neg Bel_i\varphi \quad \text{(5-}Bel_i\text{)}$$

Thereby an agent's beliefs are consistent (D-Bel_i), and an agent is aware of his beliefs (4-Bel_i) and disbeliefs (5-Bel_i).

3.3.4 Time

The operators G and H have the linear tense logic $\mathbf{S4.3}_t$ (cf. Burgess (2002)) which is a normal modal logic **K** for each operator plus the following axioms:

$$G\varphi \rightarrow \varphi \qquad \text{(T-}G\text{)}$$

$$(F\varphi \wedge F\psi) \rightarrow (F(\varphi \wedge F\psi) \vee F(\psi \wedge F\varphi)) \qquad \text{(3-}F\text{)}$$

$$G\varphi \rightarrow GG\varphi \qquad \text{(4-}G\text{)}$$

$$H\varphi \rightarrow \varphi \qquad \text{(T-}H\text{)}$$

$$(P\varphi \wedge P\psi) \rightarrow (P(\varphi \wedge P\psi) \vee P(\psi \wedge P\varphi)) \qquad \text{(3-}P\text{)}$$

$$H\varphi \rightarrow HH\varphi \qquad \text{(4-}H\text{)}$$

$$\varphi \rightarrow GP\varphi \qquad \text{(CONV-GP)}$$

$$\varphi \rightarrow HF\varphi \qquad \text{(CONV-HF)}$$

(T-G) and (T-H) mean that both future and past include the present.

(3-F) and (3-P) indicate that if two formulas are true at two instants in the future (resp. in the past) then one is necessarily true before the other. This entails that time is linear in the future and in the past (cf. Fig. 2).

(CONV-GP) and (CONV-HF) are the conversion axioms. They axiomatize that the accessibility relation for G is the converse of that for H.

3.3.5 Probability

The notion of subjective probability measure is captured here semantically by the fact that probable worlds belong to the set of believed worlds. This approach is based on neighborhood functions (as opposed to probability distributions).

The logic of $Prob$ is weaker than the logic of belief. In particular, the formula $(Prob_i\varphi \wedge Prob_i\psi) \rightarrow Prob_i(\varphi \wedge \psi)$ is not valid, and this is enough to make it a non-normal modal logic in the sense of Chellas (1980, Theorem 4.3).

The semantical conditions validate the following principles:

$$\frac{\varphi \rightarrow \psi}{Prob_i\varphi \rightarrow Prob_i\psi} \qquad \text{(RM-}Prob_i\text{)}$$

$$\frac{\varphi}{Prob_i\varphi} \qquad \text{(RN-}Prob_i\text{)}$$

$$Prob_i\varphi \rightarrow \neg Prob_i\neg\varphi \qquad \text{(D-}Prob_i\text{)}$$

3.3.6 Desirability

The logic of desirability is standard deontic logic (SDL) (Chellas 1980) and is also expressed in terms of ideal worlds: the logic associated with the operators Des_i is **KD**, i.e. the normal modal logic **K** plus the following axiom:

$$Des_i\varphi \rightarrow \neg Des_i\neg\varphi \qquad \text{(D-}Des_i\text{)}$$

which makes desirabilities consistent.

It has been argued that in principle (e.g. Castelfranchi and Paglieri (2007) and also Lang et al. (2002)), desirability is closed neither under implication nor under conjunction: It may be desirable for me to marry Ann and it may be desirable for me to marry Beth, but this does not imply that it is desirable for me to be a bigamist. Though, for the sake of simplicity, our Des_i operators are normal and hence closed under both conjunction and implication.

3.3.7 Ideals

Just as for desirability, the logic of ideality is standard deontic logic SDL, i.e. the normal modal logic **K** plus the following axiom:

$$Idl\varphi \rightarrow \neg Idl\neg\varphi \qquad \text{(D-}Idl_i\text{)}$$

which makes ideals consistent.

3.3.8 Mix axioms

The interdependencies between some modal operators are captured by the following axioms. First, the following introspection axioms express that the agents are aware of their probabilities and desirabilities:

$$Prob_i\varphi \rightarrow Bel_i\, Prob_i\varphi \qquad \text{(4-MIX1)}$$

$$\neg Prob_i\varphi \rightarrow Bel_i\neg Prob_i\varphi \qquad \text{(5-MIX1)}$$

$$Des_i\varphi \rightarrow Bel_i\, Des_i\varphi \qquad \text{(4-MIX2)}$$

$$\neg Des_i\varphi \rightarrow Bel_i\neg Des_i\varphi \qquad \text{(5-MIX2)}$$

From these axioms plus (D-Bel_i), we can easily prove their converse. For example, we deduce the converse of (4-MIX1) from $Bel_i\, Prob_i\varphi \rightarrow \neg Bel_i\neg\, Prob_i\varphi$ by (D-Bel_i), and $\neg Bel_i\neg Prob_i\varphi \rightarrow\ Prob_i\varphi$ by (5-MIX1). We therefore have the equivalences $Prob_i\varphi \leftrightarrow Bel_i\, Prob_i\varphi$ and $\neg Prob_i\varphi \leftrightarrow Bel_i\neg Prob_i\varphi$.

Then the following axioms express that actions are public:

$$Done_\alpha\top \rightarrow Bel_i\ Done_\alpha\top \qquad \text{(4-MIX3)}$$

$$\neg Done_\alpha \top \rightarrow Bel_i \neg Done_\alpha \top \quad \text{(5-MIX3)}$$

From these axioms plus (D-Bel_i) we can easily prove their converse, and we thus have the equivalences $Done_\alpha \top \leftrightarrow Bel_i\ Done_\alpha \top$ and $\neg Done_\alpha \top \leftrightarrow Bel_i \neg Done_\alpha \top$.

We axiomatize the inclusion of elements of neighborhoods in epistemic states through the following axiom:

$$(Bel_i \varphi \wedge Prob_i \psi) \rightarrow Prob_i(\varphi \wedge \psi) \quad \text{(C-MIX)}$$

which allows to derive the following theorems:

$$Bel_i \varphi \rightarrow Prob_i \varphi \quad (2)$$
$$Prob_i \varphi \rightarrow \neg Bel_i \neg \varphi \quad (3)$$

Time and action are linked: if φ is always true in the future then φ will be true after every action performance. Similarly, if φ was always true in the past, then φ was true before every performance of an action. So:

$$G\varphi \rightarrow After_\alpha \varphi \quad \text{(GA-MIX)}$$

$$H\varphi \rightarrow Before_\alpha \varphi \quad \text{(HB-MIX)}$$

Finally, desirability persists, i.e. it is preserved through time.

$$Des_i \varphi \rightarrow G Des_i \varphi \quad \text{(Pers-}Des_i\text{)}$$

$$\neg Des_i \varphi \rightarrow G \neg Des_i \varphi \quad \text{(Pers-}\neg Des_i\text{)}$$

These two principles both entail the equivalences $Des_i \varphi \leftrightarrow G Des_i \varphi$ and $\neg Des_i \varphi \leftrightarrow G \neg Des_i \varphi$.

For the same reasons, ideals also persist:

$$Idl \varphi \rightarrow G Idl \varphi \quad \text{(Pers-}Idl_i\text{)}$$

$$\neg Idl \varphi \rightarrow G \neg Idl \varphi \quad \text{(Pers-}\neg Idl_i\text{)}$$

These two principles entail that we have an equivalence.

The "no forgetting" constraint linking actions and belief is captured by the following axiom:

$$(Bel_i\ After_\alpha \varphi \wedge \neg Bel_i\ After_\alpha \bot) \rightarrow After_\alpha Bel_i \varphi \quad \text{(NF-}Bel_i\text{)}$$

This axiom expresses that the agents do not forget their previous alternatives, when the performance of the action is not surprising for them ($\neg Bel_i\ After_\alpha \bot$ reads "agent i does not believe that action α is inexecutable"). Otherwise, if $Bel_i\ After_\alpha \bot$ holds,

then the agent has to revise his beliefs upon learning that α occurred. We do not go into this here, and refer the reader to Herzig and Longin (2002).

In the next two sections we are going to put to work logic *EL*, and are going to express twenty from the twenty-two emotions of OCC theory. (We do not define the remaining emotions of love and hate because they would require a first order modal logic.) For each of these twenty emotions, we first give the informal definition of OCC theory, and then our definition in terms of logical formulas. In order to support the accuracy of our definitions, we show that they can account for the examples illustrating the emotions in Ortony et al. (1988). Below, the quoted pages all refer to this book.

4 Event-based emotions

The event-based branch of OCC theory contains emotion types whose eliciting conditions depend on the evaluation of an event with respect to the agent's goals. *Desirability* is the central variable accounting for the impact that an event has on an agent's goals, namely how it helps or impedes their achievement.

In our formalism, an event is something that may occur without any agent intending it, and is thus different from an action (that is always intentional). According to OCC theory an event can have several aspects, each of them possibly triggering a different emotion. In this paper we represent an emotion as an abbreviation of a complex formula. Moreover we assume that what Ortony et al. call the different aspects of an event can be considered as consequences of the primary event. For example the event of receiving a letter from a bailiff to inform you that you are going to inherit some money from a deceased relative has (at least) two aspects: the undesirable aspect is that your relative is dead, while the desirable aspect is that you get some money. We represent these two aspects as if they were two separate secondary events actually resulting from the primary event. While the same primary event can trigger opposite emotions (sadness that your relative died and joy of getting some money), we consider in our formalization that these emotions are attached to two different secondary events, but not to the primary event.

According to OCC theory, an event is desirable for an agent if its consequence φ is more beneficial (furthering his goals) than harmful (impeding some goals). As we said before (see Sect. 2.2), desirability depends on the agent's goals, but we do not want to enter into the details of this computation here, and assume that the agent's desirability values are given by the *Des* operators. We directly use this variable in the definitions of event-based emotions (cf. Sect. 3 for our modelling of desirability).

4.1 Well-being emotions

The emotion types in this group have eliciting conditions focused on the desirability for the self of an event. An agent feels joy (resp. distress) when he is pleased (resp. displeased) about a desirable (resp. undesirable) event.

$$Joy_i\varphi \stackrel{def}{=} Bel_i\varphi \wedge Des_i\varphi$$
$$Distress_i\varphi \stackrel{def}{=} Bel_i\varphi \wedge Des_i\neg\varphi$$

Consider an example situation from (Ortony et al. 1988, p. 88) where a man i learns that he inherits of a small amount of money (m) from a remote and unknown relative that has died (d). This is expressed by the formula $Bel_i(m \wedge d)$. Then i feels **joy** because he focuses on the desirable event ($Des_i m$) and not on the undesirable event d. This man does not feel distress about his relative's death since he did not know the relative, his death is not undesirable for him ($\neg Des_i \neg d$). On the contrary, a man j (p. 89) who runs out of gas on the freeway ($Bel_j o$) feels **distress** because this is undesirable for him ($Des_j \neg o$).

4.2 Prospect-based emotions

The emotion types in this group have eliciting conditions focused on the desirability for self of an anticipated (uncertain) event, that is actively prospected. OCC uses a local intensity variable called *likelihood*, accounting for the expected probability of the event to occur. We model likelihood by the following abbreviation $Expect_i$.

Definition 1 $Expect_i\varphi \stackrel{def}{=} Prob_i\varphi \wedge \neg Bel_i\varphi$

$Expect_i\varphi$ reads "agent i expects φ to be true but envisages that it could be false". We can notice that if i expects something then he necessarily envisages it:

$$Expect_i\varphi \rightarrow \neg Bel_i\neg\varphi \tag{4}$$

From (D-$Prob_i$) we can easily prove the consistency of expectations:

$$Expect_i\varphi \rightarrow \neg Expect_i\neg\varphi \tag{5}$$

An agent feels hope (resp. fear) if he is "pleased (resp. displeased) about the **prospect** of a desirable (resp. undesirable) event". Note that the object of hope is not necessarily about the future: I might ignore whether my email has been delivered to the addressee, and hope it has been so.

$$Hope_i\varphi \stackrel{def}{=} Expect_i\varphi \wedge Des_i\varphi$$
$$Fear_i\varphi \stackrel{def}{=} Expect_i\varphi \wedge Des_i\neg\varphi$$

The agent feels fear-confirmed (resp. satisfaction) if he is "displeased (resp. pleased) about the **confirmation** of the prospect of an undesirable (resp. desirable) event". We use here our operator P (see Definition Def$_P$, just before Sect. 3.2) to represent what was true in the past.

$$Satisfaction_i\varphi \stackrel{def}{=} Bel_i PExpect_i\varphi \wedge Des_i\varphi \wedge Bel_i\varphi$$
$$FearConfirmed_i\varphi \stackrel{def}{=} Bel_i PExpect_i\varphi \wedge Des_i\neg\varphi \wedge Bel_i\varphi$$

Given our definitions of joy and distress $Satisfaction_i\varphi$ can be written more concisely $Bel_i PExpect_i\varphi \wedge Joy_i\varphi$, and $FearConfirmed_i\varphi$ can be written $Bel_i P Expect_i \wedge Distress_i\varphi$.

The agent feels relief (resp. disappointment) if he is "pleased (resp. displeased) about the **disconfirmation** of the prospect of an undesirable (resp. desirable) event".

$$Relief_i\varphi \stackrel{def}{=} Bel_i PExpect_i\neg\varphi \wedge Des_i\varphi \wedge Bel_i\varphi$$
$$Disappointment_i\varphi \stackrel{def}{=} Bel_i PExpect_i\neg\varphi \wedge Des_i\neg\varphi \wedge Bel_i\varphi$$

For example a woman w who applies for a job (p. 111) might feel **fear** if she expects not to be offered the job ($Expect_w\neg beHired$), or feel **hope** if she expects that she will be offered it ($Expect_w beHired$). Then, if she hoped to get the job and finally gets it, she feels **satisfaction**; and if she does not get it, she feels **disappointment**. An employee e (p. 113) who expects to be fired ($Expect_e beFired$) will feel **fear** if it is undesirable for him ($Des_e\neg beFired$), but not if he already envisaged to quit this job since in this case we can suppose that this is not undesirable for him ($\neg Des_e\neg beFired$). In the first case he will feel **relief** when he is not fired ($Bel_e\neg beFired$), and **fear-confirmed** when he is.

4.3 Fortunes-of-others emotions

The emotion types in this group have eliciting conditions focused on the presumed desirability for another agent. They use three local intensity variables: *desirability for other*, *deservingness*, and *liking*. *Desirability for other* is the assessment by i of how much the event is desirable for the other one (j). *Deservingness* represents how much agent i believes that agent j deserved what occurred to him. (It often depends on *liking*, i.e. i's attitude towards j, but we cannot account for this here because we do not consider attraction emotions.)

We thus have to model these variables. First, we can represent *desirability for other* by a belief about the other's desire: $Bel_i Des_j\varphi$ reads "agent i believes that φ is desirable for agent j". Second, we represent *liking* through non-logical axioms. For example, when John likes Mary this means that if John believes that it is desirable for Mary to be rich, then it is desirable for John that Mary be rich, or better: gets to know that she is rich. In formulas: ($Bel_{john} Des_{mary} rich \rightarrow Des_{john} Bel_{mary} rich$). These axioms will be global hypotheses in deductions, in the sense that they are supposed to be known by all agents and to hold through time. Third, we simplify the concept of *deservingness* by assuming that agents want (have a goal) that any agent gets what he deserves. Then when an event (represented with the formula φ) occurs that is believed

(by i) to be deserved by j, this event will be desirable for i (we write it $Des_i Bel_j \varphi$[8]), since it achieves his goal. So finally i can desire that j believes φ either because he believes that j desires φ and j is his friend, or because he believes that j desires $\neg\varphi$ and j is his enemy, or because he believes that j deserved φ.

There are two good-will (or empathetic) emotions: an agent feels happy for (resp. sorry for) another agent if he is pleased (resp. displeased) about an event presumed to be desirable (resp. undesirable) for this agent.

$$HappyFor_{i,j}\varphi \stackrel{def}{=} Bel_i\varphi \wedge Bel_i Des_j \varphi \wedge Des_i Bel_j \varphi$$
$$SorryFor_{i,j}\varphi \stackrel{def}{=} Bel_i\varphi \wedge Bel_i Des_j \neg\varphi \wedge Des_i \neg Bel_j \varphi$$

There are two ill-will emotions: an agent feels resentment (resp. gloating) towards another agent if he is displeased (resp. pleased) about an event presumed to be desirable (resp. undesirable) for this agent.

$$Resentment_{i,j}\varphi \stackrel{def}{=} Bel_i\varphi \wedge Bel_i Des_j \varphi \wedge Des_i \neg Bel_j \varphi$$
$$Gloating_{i,j}\varphi \stackrel{def}{=} Bel_i\varphi \wedge Bel_i Des_j \neg\varphi \wedge Des_i Bel_j \varphi$$

For example (p. 95) Fred feels **happy for** Mary when he learns that she wins a thousand dollars ($Bel_f w \wedge Bel_f Bel_m w$), because he believes this is desirable for her ($Bel_f Des_m w$), and she is his friend, i.e. he *likes* her. As said above, we represent this notion of *liking* with non-logical global axioms representing one's interest in the well-being of one's friends. In this case, if Fred believes that it is desirable for Mary to win, then it is desirable for him that she gets to know that she won ($Bel_f Des_m w \rightarrow Des_f Bel_m w$).[9]

A man i (p. 95) can feel **sorry for** the victims v of a natural disaster ($Bel_i disaster \wedge Bel_i Bel_v disaster \wedge Bel_i Des_v \neg disaster$) without even knowing them, because he has an interest that people do not suffer undeservedly, so there it is undesirable for him that people suffer from this disaster ($Des_i \neg Bel_v disaster$).

An employee e (p. 99) can feel **resentment** towards a colleague c who received a large pay raise ($Bel_e pr$, $Bel_e Des_c pr$), what he believes to be desirable for this colleague ($Bel_c Des_e pr$), because he thinks this colleague is incompetent and thus does not deserve this raise. As we said above, we represent the deservingness (whatever the reason for this belief, here the incompetence) with a desirability concerning the occurrence of this event to the other agent (here $Des_e \neg Bel_c pr$) that is its consequence.

[8] The emotions types in this group result from the appraisal of an event **concerning another agent**. We represent the occurrence of this event φ to agent j by the formula $Bel_j \varphi$. Then the $Des_i Bel_j \varphi$ element of the definitions means that this event occurring to j is desirable for i, which is our way to distinguish between good-will and ill-will emotions.

[9] Note that Fred may not be happy for Mary if she was not to learn about her gain in the future. However, even if she does not know yet that she won, he can feel happy for her just because he considers it probable that she will learn it at a future moment (without being sure of that). For example, Mary may have not seen the results yet, and Fred cannot be sure that she will not forget to check them.

Finally, Nixon's political opponents (o) (p. 104) might have felt **gloating** about his departure from office ($Bel_o\ d \wedge Bel_o\ Bel_{nixon}\ d$), because they believed it to be undesirable for him ($Bel_o\ Des_{nixon} \neg d$) and they thought it was deserved (as above we identify deservingness with a desire, here: $Des_o\ Bel_{nixon}\ d$).

Remark 1 Our formalization of liking leads to the following question: what if i believes that for some reasons j will never learn that φ? In many situations it is certainly odd to say that i is happy or sorry for another agent j about something that j will never know, and thus about what j will never be happy or sad about himself. OCC theory does **not** require that j should know about this event that is important to him, and we therefore have chosen to stay as close as possible to it. However, one might wish to sharpen the definitions of these four emotions by requiring that it must be at least probable for i that j learns about the event at some time point in the future. This can be implemented by adding the further conjunct $Prob_i F Bel_j \varphi$ to our definitions of the four fortunes-of-others emotions.

5 Agent-based emotions

The agent-based branch of OCC theory contains emotion types whose eliciting conditions depend on the judgement of the praiseworthiness of an action, with respect to standards.

In our sense an action is something that is performed intentionally **(deliberately, purposely)** by an agent. It thus differs from an event. If an agent performs an action not purposely, like sneezing, we call this an event. This distinction allows to capture implicitly Lazarus' variable of attribution of responsibility that is needed for emotions like anger: an agent is always responsible for his actions.

An action is *praiseworthy* (resp. *blameworthy*) when it upholds (resp. violates) standards. The standards under concern are supposed to be internalized, i.e. the (evaluating) agent has adopted them. We express these internalized standards for agent i through the deontic operators Idl_i.

5.1 Attribution emotions

The emotion types in this group have eliciting conditions focused on the approving of an agent's action. They use two local intensity variables. *Strength of unit* intervenes in self-agent emotions to represent the degree to which the agent identifies himself with the author of the action, allowing him to feel pride or shame when he is not directly the actor; for example one can be proud of his son succeeding in a difficult examination, or of his rugby team winning the championship; in this paper we only focus on emotions felt by the agent about his own actions, because this variable is too complex to be represented in our framework. *Expectation deviation* accounts for the degree to which the performed action differs from what is usually expected from the agent, according to his social role or category.[10]

[10] In self-agent emotions, the agent refers to his stereotyped representation of himself.

We express this notion of expectation with the formula $Prob_j After_{i:\alpha} \neg\varphi$ reading "j considers it probable that after i performs α, φ is false", i.e. j expects i not to achieve φ as a result of his action, for example because it is difficult.[11] The deviation comes from the fact that after the execution of α, j believes that φ is nevertheless true, contrarily to what he expected.[12] This prevents the agent from feeling attribution emotions too often. Indeed, we often respect the law without being proud, and we often violate standards without being ashamed. Therefore we consider that the standards have to be internalized and accepted by the agent as belonging to his values. This allows an agent to feel no emotion, even concerning an (un)ideal action, when this is not important for him. For example someone who likes to wear strange (unideal) clothes would not feel ashamed about this if it is what he desires to wear, but would feel so if he was forced to wear such clothes.

Finally, we do not impose that the ideal was conscious at the moment of the action. For example one can feel shame about having performed an action when one realizes that it was blameworthy, even if one ignored that at the time when the action was performed. Ideally, we should not impose it either for probability, but the $Prob_i$ operators are intrinsically epistemic (i.e. semantically, probable worlds are a subset of possible worlds compatible with the agent's beliefs); so technically it is difficult to do so.

In the sequel, $Emotion_i(i{:}\alpha, \varphi)$ (resp. $Emotion_{i,j}(j{:}\alpha, \varphi)$) abbreviates $Emotion_i(Done_{i:\alpha}\top, \varphi)$ (resp. $Emotion_{i,j}(Done_{j:\alpha}\top, \varphi)$) where $Emotion$ is the name of an emotion.

Remark 2 These emotions are about an action α that the agent believes to have influenced the proposition φ: the agent believes that "if he had not performed action α, φ would probably be false now". Though, our language is not expressive enough to represent this counterfactual reasoning, so we make the hypothesis that the agent i believes that α and φ are linked in this way. The following emotions do make sense only when this is the case.

Self-agent emotions: an agent feels pride (resp. shame) if he is approving (resp. disapproving) of his own praiseworthy (resp. blameworthy) action.

$$Pride_i(i{:}\alpha, \varphi) \stackrel{def}{=} Bel_i\ Done_{i:\alpha}(Idl_i\ Happens_{i:\alpha}\varphi \wedge Prob_i\ After_{i:\alpha}\neg\varphi) \wedge Bel_i\varphi$$

$$Shame_i(i{:}\alpha, \varphi) \stackrel{def}{=} Bel_i\ Done_{i:\alpha}(Idl_i \neg Happens_{i:\alpha}\varphi \wedge Prob_i\ After_{i:\alpha}\neg\varphi) \wedge Bel_i\varphi$$

[11] In the following, whatever the cause of the unexpectedness is (for example difficulty), we only formalize the consequence (the unexpectedness itself) with the above formula.

[12] What is unexpected is not only the performance of the action but also its result φ; actually, when the result is not important, φ is $\top$ and then it is the very performance of the action that is unexpected. For example it is unexpected from a wise child to steal something in a shop, whatever the result of his action is (did he succeed or not), so we write: $Prob_i\ After_{child:steal}\bot$.

Emotions involving another agent:[13] an agent feels admiration (resp. reproach) towards another agent if he is approving (resp. disapproving) of this agent's praiseworthy (resp. blameworthy) action.

$$Admiration_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Bel_i\, Done_{j:\alpha}(Idl_i\ Happens_{j:\alpha}\varphi \wedge Prob_i\ After_{j:\alpha}\neg\varphi) \wedge Bel_i\varphi$$

$$Reproach_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Bel_i\ Done_{j:\alpha}(Idl_i\neg Happens_{j:\alpha}\varphi \wedge Prob_i\ After_{j:\alpha}\neg\varphi) \wedge Bel_i\varphi$$

For example, a woman m feels **pride** (p. 137) of having saved the life of a drowning child because she performed the action α (to jump into the water to try to save him) with the successful result s (the child is safe): $Bel_m\ Done_{m:\alpha}\top \wedge Bel_m s$.[14] Moreover she now believes that before the action, it was ideal to save the child and she internalized this ideal ($Idl_m\ Happens_{m:\alpha}\top$), but she had not much chances to succeed:[15] $Prob_m\ After_{m:\alpha}\neg s$.

A rich elegant lady l (p. 142) would feel **shame** when caught while stealing clothes in an exclusive boutique ($Shame_l(\alpha, \top)$), where α is the action to steal, because she has performed an action that was unideal for her[16] ($Idl_l\neg Happens_{l:\alpha}\top$) and improbable to be performed by her ($Prob_l\ After_{l:\alpha}\bot$) due to her social role. The result of the action is $\top$ here because this emotion does not depend on the success or failure of the action but on its very performance.

A physicist p's colleagues c (p. 145) feel **admiration** towards him for his Nobel-prize-winning work ($Bel_c Done_{p:\alpha}\top \wedge Bel_c w$, where α is the action of conducting experiments, with the result w of obtaining Nobel-prize-deserving findings) because they internalized this result as ideal.[17] ($Idl_c\ Happens_{p:\alpha} w$) and difficult thus unexpected ($Prob_c\ After_{p:\alpha}\neg w$). As we said above, the difficulty of an action is one possible reason for its result to be unexpected. Here to simplify we do not formalize the very notion of difficulty but only its consequence, i.e. the unexpectedness of the result, which is what we are interested in when we define attribution emotions.

A man i may feel **reproach** towards a driver j (p. 145) who drives without a valid license ($Bel_i\ Done_{j:\delta}\top$, where δ is the action to drive without a valid license), because

[13] When $i = j$, these emotions correspond to the self-agent emotions (cf. Theorem 5).

[14] Actually, she also believes that she influenced this result by her action, i.e. she believes that if she had not jumped into the water the child could have drowned; as we said it before, we cannot express this causal link in our language, so our account is incomplete in that respect.

[15] Thus, she would not feel pride after saving the child if she believes it was easy for her.

[16] Actually actions do not obligatorily follow moral values. The lady may have been driven by the desire to possess the object, violating her ideals. But this example seems to be a borderline case, since she could have bought the object instead.

[17] Here, what is ideal is not only the execution of the action but its execution with this result. Similarly, in the case of negative emotions, what is unideal is not the happening of the action, but its happening with a given result: $Idl_i\neg Happens_{i:\alpha}\varphi$. This is compatible with the fact that the action itself could be ideal: $Idl_i\ Happens_{i:\alpha}\top$. For example, it is ideal to participate, but unideal to lose when you are expected to win.

it is forbidden and he considers this obligation to be important ($Idl_i \neg Happens_{j:\delta} \top$) and unexpected from a driver ($Prob_i After_{j:\delta} \bot$).

5.2 Compound emotions

These emotions occur when the agent appraises both the consequences of the event and its agency. They are thus the result of a combination of attribution emotions about an action α with result φ, and well-being emotions about this result φ.

$$Gratification_i(i{:}\alpha, \varphi) \stackrel{def}{=} Pride_i(i{:}\alpha, \varphi) \wedge Joy_i\varphi$$
$$Remorse_i(i{:}\alpha, \varphi) \stackrel{def}{=} Shame_i(i{:}\alpha, \varphi) \wedge Distress_i\varphi$$
$$Gratitude_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Admiration_{i,j}(j{:}\alpha, \varphi) \wedge Joy_i\varphi$$
$$Anger_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Reproach_{i,j}(j{:}\alpha, \varphi) \wedge Distress_i\varphi$$

For example, a woman i may feel **gratitude** (p. 148) towards the stranger j who saved her child from drowning ($Bel_i\ Done_{j:\alpha} \top \wedge Bel_i s$, where $j{:}\alpha$ is j's action to jump in the water, and s is the result: her child is safe). Indeed, i feels admiration towards j because of j's ideal but difficult (i.e. before it, $Prob_i\ After_{j:\alpha} \neg s$ held) action. Moreover the result of j's action ($Bel_i s$) is desirable for i ($Des_i s$), so i also feels joy about it ($Joy_i s$).

Similarly, a woman w (p. 148) may feel **anger** towards her husband h who forgets to buy the groceries ($Bel_w Done_{h:\alpha} \top$, where α is his action to go shopping, and $Bel_w \neg g$, where g reads "there are groceries for dinner"), because w reproaches this unideal result to h (it was not the expected result of the action: $Prob_w After_{h:\alpha} g$), and she is also sad about it ($Distress_w \neg g$) because she desired to eat vegetables ($Des_w g$).

The physicist p may feel **gratification** about winning the Nobel prize because he performed a successful execution of action α (performing experiments), achieving the ideal result n (he receives the Nobel prize), and thus feels pride; and this result is not only socially ideal but also desirable for him[18] ($Des_p n$), so pride combines with joy.

Finally, a spy may feel **remorse** (p. 148) about having betrayed his country (action ω) if he moreover caused undesirable damages (result d): $Shame_{spy}(\omega, d) \wedge Distress_{spy} d$.

6 Formal properties

In the previous section we started from OCC theory, extracted its key concepts, and casted them into logical definitions of twenty emotions. The first benefit of our work is to disambiguate these definitions, that might be debatable when expressed in natural language. For example, we had to decide between two different options in the case of fortunes-of-others emotions, depending on whether we accepted that one can feel

[18] This is not always true. For example, a child may personally desire not to go to school, while it is ideal to go.

happy about some good news for somebody else even if we believe that this person will never learn the good news. Furthermore, the use of logic enables to reason about the formalized concepts and to derive properties. In contrast, properties of emotions are always debatable when defined informally, and many debates have occurred as research progressed.

In this section we expose some theorems following from our definitions, mainly concerning the causal and temporal links that emotions have with each other. These theorems are consistent with OCC theory; sometimes they even go beyond it, but they always remain intuitive. Moreover, what is interesting is that the formal proofs of these theorems make that they are not debatable on their own once one has accepted the principles of the logic. This shows again the advantages of formal reasoning about emotions. The reader who is interested in the proofs of these theorems is referred to the appendix.

6.1 Prospect-based emotions and their confirmation

If an agent remembers that at a moment in the past he was feeling a prospect-based emotion about φ, and if he now knows whether φ is true or false, then it follows by the laws of our logic that he feels the corresponding confirmation emotion.

Theorem 1 (Temporal link from prospect to confirmation).

$$\begin{aligned} &\vdash (Bel_i\, PHope_i\varphi \wedge (Bel_i\varphi \vee Bel_i\neg\varphi)) \\ &\rightarrow Satisfaction_i\varphi \vee Disappointment_i\neg\varphi \end{aligned} \tag{a}$$

$$\begin{aligned} &\vdash (Bel_i\, PFear_i\varphi \wedge (Bel_i\varphi \vee Bel_i\neg\varphi)) \\ &\rightarrow Relief_i\varphi \vee FearConfirmed_i\neg\varphi \end{aligned} \tag{b}$$

Moreover, we can prove that an agent cannot feel simultaneously two emotions concerning the confirmation and the disconfirmation of the same expectation.

Theorem 2 (Inconsistency between confirmation and disconfirmation).

$$\vdash \neg(Satisfaction_i\varphi \wedge Disappointment_i\neg\varphi) \tag{a}$$

$$\vdash \neg(FearConfirmed_i\varphi \wedge Relief_i\neg\varphi) \tag{b}$$

The proof follows from the rationality axiom for belief.

Please note that on the contrary, we cannot prove inconsistencies between relief and satisfaction, or between fear-confirmed and disappointment. This is because $Bel_i\, PExpect_i\neg\varphi$ and $Bel_i\, PExpect_i\varphi$ are consistent, i.e. the agent may have expected φ at one moment in the past and $\neg\varphi$ at another moment.[19] We can only prove that

[19] Thus, our current definitions of confirmation and disconfirmation emotions may not be precise enough to entail this intuitive inconsistency. Actually in linear temporal logic with $Until$ and $Since$ operators, we could write for example $Relief_i\varphi \stackrel{def}{=} Bel_i\, P(\neg Expect_i\neg\varphi Since\ Expect_i\varphi) \wedge Des_i\varphi \wedge Bel_i\varphi$.

these two expectations $Expect_i \neg\varphi$ and $Bel_i\, PExpect_i \varphi$ cannot occur at the same time. (This is the theorem (5) of Sect. 4.2).

We can prove that the positive confirmation emotions imply joy, and that the negative confirmation emotions imply distress. This is intuitive, and in agreement with Ortony et al. 's definitions.

Theorem 3 (Link between confirmation and well-being emotions).

$$\vdash Satisfaction_i \varphi \rightarrow Joy_i \varphi \quad \text{(a)}$$

$$\vdash FearConfirmed_i \varphi \rightarrow Distress_i \varphi \quad \text{(b)}$$

$$\vdash Relief_i \varphi \rightarrow Joy_i \varphi \quad \text{(c)}$$

$$\vdash Disappointment_i \varphi \rightarrow Distress_i \varphi \quad \text{(d)}$$

6.2 Fortunes-of-others emotions

In this paragraph we will ground on reinforced definitions of fortunes-of-others emotions, that we denote them by $Emotion'_{i,j}\varphi$ where $Emotion'$ ranges over the four fortunes-of-other emotions in $\{HappyFor, SorryFor, Resentment, Gloating\}$. These reinforced definitions are obtained from our definitions by adding the further conjunct $Prob_i FBel_j \varphi$ to them, as we suggested in Remark 1. For instance $HappyFor'_{i,j}\varphi \stackrel{def}{=} HappyFor'_{i,j}\varphi \wedge Prob_i FBel_j \varphi$.

We can prove that if the agent i feels a fortune-of-other emotion towards another agent j about φ, then it is at least probable for i that j is going to feel the corresponding well-being emotion about φ at some moment in the future.

This leads us to believe that OCC definitions of these emotions may be to vague, since they do not allow to deduce these properties while they are quite intuitive.

Theorem 4 (From fortune-of-other emotion to image of other).

$$\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i FJoy_j \varphi \quad \text{(a)}$$

$$\vdash SorryFor'_{i,j}\varphi \rightarrow Prob_i FDistress_j \varphi \quad \text{(b)}$$

$$\vdash Resentment'_{i,j}\varphi \rightarrow Prob_i FJoy_j \varphi \quad \text{(c)}$$

$$\vdash Gloating'_{i,j}\varphi \rightarrow Prob_i FDistress_j \varphi \quad \text{(d)}$$

If an agent i feels a fortune-of-other emotion towards another agent about φ, and i is not sure that j will learn about the event φ, then i feels a corresponding prospect-based emotion about j believing φ.

Theorem 5 (Consequences of fortunes-of-others emotions).

$$\vdash (HappyFor'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi) \rightarrow Hope_i FBel_j\varphi \quad \text{(a)}$$

$$\vdash (SorryFor'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi) \rightarrow Fear_i FBel_j\varphi \quad \text{(b)}$$

$$\vdash (Resentment'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi) \rightarrow Fear_i FBel_j\varphi \quad \text{(c)}$$

$$\vdash (Gloating'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi) \rightarrow Hope_i FBel_j\varphi \quad \text{(d)}$$

6.3 Links between self-agent and other-agent attribution emotions

We can prove that an other-agent emotion towards oneself is equivalent to the corresponding self-agent emotion. This is rather intuitive, all the more Ortony et al. introduce the term *self-reproach* for shame.

Theorem 6 (Other-agent emotions towards oneself).

$$\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Pride_i(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\vdash Reproach_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Shame_i(i{:}\alpha, \varphi) \quad \text{(b)}$$

We can prove that if another agent j feels an attribution emotion towards an agent i about a given action with a given result, then the agent i does not inevitably feel the corresponding self-agent attribution emotion. That is, one can admire you about a given action while you are not proud about it.

Theorem 7 (Other-agent emotion does not force self-agent emotion).

$$\nvdash Bel_i Admiration_{j,i}(i{:}\alpha, \varphi) \rightarrow Pride_i(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\nvdash Bel_i Reproach_{j,i}(i{:}\alpha, \varphi) \rightarrow Shame_i(i{:}\alpha, \varphi) \quad \text{(b)}$$

Both prospect-based emotions and attribution emotions involve probabilities. We thus get interested in their temporal links with each other. We can prove that if an agent feels an attribution emotion about an action with a given result, and that before this action he envisaged that it could happen with this result and had a corresponding desire, then at this moment he felt a prospect-based emotion about the performance of this action with this result (namely about the success or failure of the action with respect to the prospected result). We have the same theorem if the agent feeling the emotion is different from the agent performing the action.

Theorem 8 (Link between prospect and attribution emotions).

$$\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}((\neg Bel_i \neg Happens_{i:\alpha}\varphi \wedge Des_i\ Happens_{i:\alpha}\varphi) \rightarrow Fear_i \neg Happens_{i:\alpha}\varphi) \quad \text{(a)}$$

$$\vdash Shame_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}((\neg Bel_i \neg Happens_{i:\alpha}\varphi \wedge Des_i \neg Happens_{i:\alpha}\varphi) \rightarrow Hope_i \neg Happens_{i:\alpha}\varphi) \quad \text{(b)}$$

$$\vdash Admiration_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i Done_{j:\alpha}((\neg Bel_i \neg Happens_{j:\alpha}\varphi \wedge Des_i\ Happens_{j:\alpha}\varphi) \rightarrow Fear_i \neg Happens_{j:\alpha}\varphi) \quad \text{(c)}$$

$$\vdash Reproach_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{j:\alpha}((\neg Bel_i \neg Happens_{j:\alpha}\varphi \wedge Des_i \neg Happens_{j:\alpha}\varphi) \rightarrow Hope_i \neg Happens_{j:\alpha}\varphi) \quad \text{(d)}$$

We can notice that we have to impose that the agent had a corresponding desire in order to make him feel fear or hope. Moral values are not sufficient to trigger these emotions, since they can be inconsistent with desires. For example one can desire to kill someone he hates while his moral values tell him not to do so.

We can also prove a kind of converse of this theorem: if the agent fears (resp. hopes) that he does not perform the action α with result φ, and that this performance is ideal for him (resp. unideal), then after he performed α, if he believes that φ is true then he feels pride (resp. shame). Actually, the agent was afraid to fail (resp. he hoped to succeed). For example someone who passes an examination and has few chances to succeed would feel afraid of failing, and then if he succeeds he would feel pride because it was difficult.

Theorem 9 (Link between attribution and prospect emotions). *If α is an action that the agent i believes to influence the proposition φ (cf. Remark 2), then:*

$$\vdash Fear_i \neg Happens_{i:\alpha}\varphi \wedge Bel_i Idl_i Happens_{i:\alpha}\varphi \rightarrow After_{i:\alpha}(Bel_i\varphi \rightarrow Pride_i(i{:}\alpha, \varphi)) \quad \text{(a)}$$

$$\vdash Hope_i \neg Happens_{i:\alpha}\varphi \wedge Bel_i Idl_i \neg Happens_{i:\alpha}\varphi \rightarrow Afteri{:}\alpha(Bel_i\varphi \rightarrow Shame_i(i{:}\alpha, \varphi)) \quad \text{(b)}$$

$$\vdash Fear_i \neg Happens_{j:\alpha}\varphi \wedge Bel_i Idl_j \neg Happens_{j:\alpha}\varphi \rightarrow After_{j:\alpha}(Bel_i\varphi \rightarrow Admiration_{i,j}(j{:}\alpha, \varphi)) \quad \text{(c)}$$

$$\vdash Hope_i \neg Happens_{j:\alpha}\varphi \wedge Bel_i Idl_j \neg Happens_{j:\alpha}\varphi \rightarrow After_{j:\alpha}(Bel_i\varphi \rightarrow Reproach_{i,j}(j{:}\alpha, \varphi)) \quad \text{(d)}$$

6.4 Inconsistencies between some emotions

We can prove several inconsistencies between pairs of emotions.

First, we can prove the inconsistency between opposite emotions about the same proposition (polar opposites), i.e. between the positive and the negative emotion of the same group. This is in agreement with the psychological definitions.

Theorem 10 (Polar inconsistencies).

$$\vdash \neg(Joy_i\varphi \wedge Fear_i\varphi)$$
$$\vdash \neg(Hope_i\varphi \wedge Fear_i\varphi)$$
$$\vdash \neg(Satisfaction_i\varphi \wedge FearConfirmed_i\varphi)$$
$$\vdash \neg(Relief_i\varphi \wedge Disappointment_i\varphi)$$
$$\vdash \neg(HappyFor'_{i,j}\varphi \wedge SorryFor_{i,j}\varphi)$$
$$\vdash \neg(Resentment_{i,j}\varphi \wedge Gloating_{i,j}\varphi)$$
$$\vdash \neg(Pride_i(i{:}\alpha, \varphi) \wedge Shame_i(i{:}\alpha, \varphi))$$
$$\vdash \neg(Admiration_{i,j}(j{:}\alpha, \varphi) \wedge Reproach_{i,j}(j{:}\alpha, \varphi))$$
$$\vdash \neg(Gratification_i(i{:}\alpha, \varphi) \wedge Remorse_i(i{:}\alpha, \varphi))$$
$$\vdash \neg(Gratitude_{i,j}(j{:}\alpha, \varphi) \wedge Anger_{ij}(j{:}\alpha, \varphi))$$

This follows in particular from the rationality axioms (D) for our operators Bel_i, Des_i, $Prob_i$ and Idl_i.

Please notice that we can still capture mixed emotions about a given event, because these mixed emotions actually concern different aspects of this event, that we represent with different formulas as if they were different consequences of the main event. For example one who loses a friend who suffered from a long and painful disease will feel sadness about the loss of his friend, and at the same time relief about the end of his friend's suffering. We thus consider that there are two appraised events: the loss of a friend, that is undesirable, and the end of his suffering, that is desirable. The initial event (the death of a friend) thus triggers a positive and a negative emotion.

Due to the properties of our probability operator, hope is not only inconsistent with fear about the same φ but also with fear about $\neg\varphi$. Actually, depending on which one is more probable between φ and $\neg\varphi$, the agent feels either hope or fear. Thus these two emotions cannot occur simultaneously.

Theorem 11 (Non simultaneity of hope and fear).

$$\vdash \neg(Hope_i\varphi \wedge Fear_i\neg\varphi)$$

This is because by definitions $Hope_i\varphi$ implies $Prob_i\varphi$ while $Fear_i\neg\varphi$ implies $Prob_i\neg\varphi$, which cannot simultaneously be the case due to the consistency of expectations (property 5 of Sect. 4.2).

Moreover, an agent cannot feel simultaneously a good-will and an ill-will emotion towards the same agent about the same issue.

Theorem 12 (Inconsistency between good-will and ill-will emotions).

$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Resentment_{i,j}\varphi) \quad (a)$$

$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad (b)$$

$$\vdash \neg(HappyFor_{i,j}\varphi \wedge Gloating_{i,j}\varphi) \quad (c)$$

$$\vdash \neg(SorryFor_{i,j}\varphi \wedge Resentment_{ij}\varphi) \quad \text{(d)}$$

The proof follows from the rationality axioms for Bel_i and Des_i (see the appendix for details).

6.5 Other interesting properties

Our formalism allows us to prove that an agent is aware of his emotions.

Theorem 13 (Emotional awareness). *For every emotion $Emotion_i$ among the twenty emotions that we have defined:*

$$\vdash Emotion_i\varphi \leftrightarrow Bel_i Emotion_i\varphi \quad \text{(a)}$$

$$\vdash \neg Emotion_i\varphi \leftrightarrow Bel_i \neg Emotion_i\varphi \quad \text{(b)}$$

This follows in particular from the introspection axioms for our operators Bel_i, $Prob_i$ and $Expect_i$.

According to Lazarus (1991), only situations that are relevant to the individual's well-being can trigger an emotion. If we consider that an event is relevant to i's well-being when it involves one of i's desires or values, then this is in agreement with the following theorem. Indeed, if the agent has no desire or ideal at all then no event is relevant to him, and thus no situation can trigger an emotion. Besides, desires and moral values are part of what Lazarus calls "ego-involvement".

Theorem 14 (Emotions and ego-involvement). *An agent who has neither desires nor ideals cannot feel any emotion.*

The proof trivially follows from the definitions of emotions, that all necessarily entail either a desire (for the event-based ones) or an ideal (for the agent-based ones). Compound emotions entail both a desire and an ideal.

7 Discussion

Logical approaches of emotions are still quite rare. J.J. Meyer is one of the few researchers to have contributed to this field. In particular he has recently proposed an approach (Meyer 2006) where emotions are considered as kinds of events, and where definitions like those presented above are the necessary conditions of the triggering of these events (that Searle would call "mental events" (Searle 1983, Chap. 3)). This model is a very interesting alternative to ours, independently from the details of the definitions respectively chosen in each approach.

Meyer (2004) proposes a logical model of emotions based on KARO, his logic of action, belief and choice (cf. van der Hoek et al. (1997) or Meyer et al. (2001)). He uses this logic to write generation rules for four emotions: joy, sadness, anger and fear, depending on the agent's plans. First, the generation conditions of these emotions only depend on the satisfaction of the agent's plans, making this model task-oriented.

Indeed, Meyer's aim, as he states himself[20], is not to be faithful to psychological definitions but to design artificial agents. On the contrary, in our work, we try to stay as close as possible to the original psychological definitions of the emotions that we formalize, through building on one of the most widely used approaches, namely Ortony, Clore, and Collins' typology. Second, this approach focuses on the individual aspects of emotions: as there is no operator to represent social standards, no social emotion like pride or shame can be represented. Finally, we thus provide an emotional formalism that is richer (with twenty emotions) and more faithful to psychology. However, our formalism is still limited to the triggering of emotions, whereas Meyer and colleagues already formalized the influence of emotions on the agents' plans (Dastani and Meyer 2006).

We would now like to highlight the assets and limitations of our own model from several points of view, namely computer science, logic and psychology.

Our model undoubtedly suffers from some limitations. First, from the logical point of view our framework lacks some expressivity. In particular we preferred not to use the full collection of existing temporal operators like *Since* or *Until* in order to keep our logic simple (see Footnote 19). As for our other choices, they are mainly due to the state of the art in BDI-like logics. First, in some places we had to approximate concepts. Most importantly, our account is incomplete as to the link between action and consequences, because propositional dynamic logic does not provide it (see Remark 2). Our logic therefore does not fully account for the notion of responsibility in agent-based emotions. Second, in some places we had to ignore concepts entirely because there is no logical operator in the literature that would allow to take them into account. Most importantly, the concept of goal does not appear in our logic, and in consequence its link with desirability is neglected. The reason here is that there exists no consensical logical analysis up to now.

Our emotions have no intensity degrees because it is not easy to design a semantics for graded operators and their evolution in time. We plan to further investigate this based on the logic of graded belief of Laverny and Lang (2005). However, despite all these limitations we believe that our formalism is expressive enough to give satisfying definitions of twenty emotions.

A point that is related to the previous one is that from the psychological point of view there are still several insufficiencies in our model. Mainly, our emotions are not quantitative: they have no intensity degree. This prevents us from fine-grained differentiations among emotions of the same type (for example: irritation, anger, rage). A second (and linked) shortcoming is that we do not manage the emotional dynamics: our emotions are persistent as long as their conditions stay true. Thereby some emotions (like Joy or $Satisfaction$) can persist *ad vitam eternam*, which is not intuitive at all. Indeed it has been established in psychology that after an emotion is triggered, its intensity decreases, and disappears when it is below a threshold. Finally, we cannot manage emotional blending of several emotions that are simultaneously triggered;

[20] "Instead of trying to capture the informal psychological descriptions exactly (or as exact as possible), we primarily look here at a description that makes sense for artificial agents." (Meyer, 2004, p.11)

Gershenson (1999) proposes an original solution to this issue. We leave these problems for further work.

Moreover we only provided a formalization of the OCC theory, that is far from being as popular in psychology as it is in computer science. It was necessary to choose one theory to begin with, but we believe that our BDI framework is expressive enough to formalize other psychological theories, all the more they often share the same appraisal variables. We already saw that we capture implicitly the control variable defined by Lazarus (1991). On the contrary we do not capture the coping potential variable because it does not intervene in the triggering of OCC emotions; however we can represent it and we did so when formalizing coping strategies Adam and Longin (2007). In this paper we only formalized the triggering of emotions, but this is the necessary starting point before formalizing their influence on any cognitive process. We neither formalize the subsequent life of emotions: their temporal decay (since we have no associated intensity degrees), and their interaction with mood or personality, but this is an interesting extension of this work.

From the logical point of view our model offers a clear semantics, which we think is quite rare in existing logical models of emotions. It also allows to highlight the power of BDI logics to reason about and disambiguate complex concepts, since we were able to prove some intuitive properties of emotions thanks to our logical definitions. It is only a logical formalism can give such unequivocal results about phenomena that are not always clearly analyzed in the psychological literature. Finally our model somehow validates BDI logics, that were designed to formalize mental attitudes, since it demonstrates that they are expressive enough to characterize as complex mental attitudes as emotions (we recall that philosophers like Searle consider emotions as complex mental attitudes, see Sect. 2.1).

From the psychologist's point of view, it could be interesting to validate theories thanks to the reasoning services offered by logic. Another way for them to take advantage from such a model is to conduct experiments with emotional agents endowed with it, instead of humans that are not always able to clearly analyze their own emotions. We have implemented such an emotional agent and have experimented it with human users analyzing the believability of its emotions. We then used the results of this experiments to derive conclusions about the underlying OCC theory. We have not conducted this work in collaboration with psychologists, but plan to do so in the near future.

Finally, from the computer science point of view this cross-disciplinary work brings an interesting contribution, since it fills the gap between psychology and the agent community. We designed a domain-independent model, based on a standard formalism, BDI logics, that are already used in a great number of agent architectures. Our model is thus ready to be implemented in any BDI agent, whatever its application may be. It will save designers the long and costly (though necessary) process of formalization of a psychological theory. Moreover it offers them a rich set of emotions that will make their agents very expressive. To illustrate all these assets we have ourselves implemented such an agent endowed with our model, only making some concessions, for example in order to add intensity degrees to emotions.

8 Conclusion

In this paper, we have formalized twenty emotions from OCC theory (all but the object-based branch), thus providing a very rich set of emotions. Moreover we have shown the soundness of our framework by illustrating each definition by an example from Ortony et al. 's book. We managed to formalize nearly the whole OCC theory with our logic, supporting the author's assumption that their theory is computationally tractable. On the contrary some appraisal variables from other theories, like Lazarus' ego-involvement, seem to be much more ambiguous and difficult to formalize.

We have privileged richness, genericity, and fidelity to the definitions over tractability. An optimization would have needed important concessions. For example Parunak et al. (2006) propose a numerical model of emotions in combat games, efficient in big real-time multi-agent systems, but which is domain-dependent.

In other works we have explored some extensions of this model. First we have provided an account of the influence of emotions on the agent's behavior by formalizing in the same BDI framework some coping strategies. According to psychologists Lazarus and Folkman (1984), appraisal and coping are indivisible. However, the formalization of each process was a full-fledged work and we thus investigated them in separate papers. Our formalization of coping strategies Adam and Longin (2007) allows to explain how an agent modifies his beliefs or intentions depending on his current emotion.

Second, we have implemented our logical model of both appraisal and coping in a BDI agent Adam (2007). This agent answers emotionally to stimuli sent by the user through the interface. This work is still in progress to implement other kinds of influence that emotions have on the agent: interaction with personality, modification of the reasoning strategies (in the sense of Forgas (1995)), impact on the agent's centers of attention (in the sense of the activation notion of Anderson and Lebiere (1998))…

Our future research will be oriented towards several different aims. First we would like to use this logical framework to formalize various existing psychological theories of emotions. Once expressed in the same language, we would be able to compare these theories. Second we would like to conduct new experiments with our BDI agents, but this time in cooperation with psychologists who could help us interpreting the results.

From a logical perspective we will further investigate the links between mental attitudes, in particular how desirability can be computed from goals. Moreover, our work currently excludes object-based emotions: in future work a modal predicate logic could allow to characterize the properties of objects and thus define the emotions triggered by their appraisal. Finally, we might unify the formalization of events and actions by moving from dynamic-logic actions to theories of agency such as STIT theory or the logic of 'brining-it-about'.

To conclude, our cross-disciplinary approach combines the advantages of logic and computational models with the expertise of psychology of emotions. Even if the resulting computational model of emotions still suffers from some limitations, we hope that it already brings some interesting contributions for computer science and logic as well as for psychology itself.

Acknowledgements We would like to thank the 3 reviewers for their thorough comments, which helped to improve the paper in several places.

9 Appendix

In order not to overload the paper, we gather in this appendix the proofs of the theorems given in the main part. This appendix is intended to help the reviewer to understand the theorems. It may be dropped in the final version of the paper.

In the proofs, $\mathcal{PL}$ refers to the Propositional Logic, and $\mathcal{ML}$ refers to the principles of normal modal logic.

Theorem 1 (Temporal link from prospect to confirmation.)

$$\vdash Bel_i\, PHope_i\varphi \wedge (Bel_i\varphi \vee Bel_i\neg\varphi) \rightarrow Satisfaction_i\varphi \vee Disappointment_i\neg\varphi \quad \text{(a)}$$

$$\vdash Bel_i\, PFear_i\varphi \wedge (Bel_i\varphi \vee Bel_i\neg\varphi) \rightarrow Relief_i\varphi \vee FearConfirmed_i\neg\varphi \quad \text{(b)}$$

To prove the Theorem 1 we need the following lemma.

Lemma 1 $\vdash Bel_i\, PDes_i\varphi \rightarrow Des_i\varphi$.

Proof 1 (of Lemma 1).

1. $\vdash Des_i\varphi \rightarrow GDes_i\varphi$ (from (Pers-Des_i))
2. $\vdash PDes_i\varphi \rightarrow PGDes_i\varphi$ (from 1. by $\mathcal{ML}$)
3. $\vdash PGDes_i\varphi \rightarrow Des_i\varphi$ (from (CONV-HF) by $\mathcal{PL}$)
4. $\vdash PDes_i\varphi \rightarrow Des_i\varphi$ (from 2. and 3. by $\mathcal{PL}$)
5. $\vdash Bel_i\, PDes_i\varphi \rightarrow Bel_i\, Des_i\varphi$ (from 4. by (RM-$\Box$) for Bel_i)
6. $\vdash Bel_i\, Des_i\varphi \rightarrow Des_i\varphi$ (from (5-MIX2) and (D-Bel_i))
7. $\vdash Bel_i\, PDes_i\varphi \rightarrow Des_i\varphi$ (from 5. and 6. by $\mathcal{PL}$)

□

Proof 2 (of Theorem 1). Case of (a). Actually it suffices to prove that (i) $Bel_i\, PHope_i\varphi \wedge Bel_i\varphi \rightarrow Satisfaction_i\varphi$ and (ii) $Bel_i\, PHope_i\varphi \wedge Bel_i\neg\varphi \rightarrow Disappointment_i\neg\varphi$ are theorems. Case of (i).

1. $\vdash Bel_i\, PHope_i\varphi \rightarrow Bel_i\, P(Expect_i\varphi \wedge Des_i\varphi)$ (from definition 1)
2. $\vdash Bel_i\, PHope_i\varphi \rightarrow Bel_i\, PExpect_i\varphi \wedge Bel_i\, PDes_i\varphi$ (by $\mathcal{ML}$)
3. $\vdash Bel_i\, PHope_i\varphi \rightarrow Bel_i\, PExpect_i\varphi \wedge Des_i\varphi$ (by Lemma 1)
4. $\vdash Bel_i\, PHope_i\varphi \wedge Bel_i\varphi \rightarrow Bel_i\, PExpect_i\varphi \wedge Des_i\varphi \wedge Bel_i\varphi$ (by $\mathcal{PL}$)
5. $\vdash Bel_i\, PHope_i\varphi \wedge Bel_i\varphi \rightarrow Satisfaction_i\varphi$ (by def. of $Satisfaction$)

We demonstrate (ii) in the same way. Case of (b): the proof is similar. □

Theorem 2 (Link between confirmation and well-being emotions)

$$\vdash Satisfaction_i\varphi \rightarrow Joy_i\varphi \quad \text{(a)}$$

$$\vdash FearConfirmed_i\varphi \rightarrow Distress_i\varphi \quad \text{(b)}$$

$$\vdash Relief_i\varphi \rightarrow Joy_i\varphi \quad \text{(c)}$$

$$\vdash Disappointment_i\varphi \rightarrow Distress_i\varphi \quad \text{(d)}$$

*Proof 3 (*of Theorem 2*).* Case of (a).

1. $\vdash Satisfaction_i\varphi \rightarrow Bel_i\varphi \wedge Des_i\varphi$ (from def. of $Satisfaction$)
2. $\vdash Satisfaction_i\varphi \rightarrow Joy_i\varphi$ (by definition of Joy)

The proof is similar for cases (b) to (d). □

Theorem 3 (From fortune-of-other emotion to image of other)

$$\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_iFJoy_j\varphi \quad \text{(a)}$$

$$\vdash SorryFor'_{i,j}\varphi \rightarrow Prob_iFDistress_j\varphi \quad \text{(b)}$$

$$\vdash Resentment'_{i,j}\varphi \rightarrow Prob_iFJoy_j\varphi \quad \text{(c)}$$

$$\vdash Gloating'_{i,j}\varphi \rightarrow Prob_iFDistress_j\varphi \quad \text{(d)}$$

Proof 4 (of Theorem 3). Case of (a).

1. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_iFBel_j\varphi \wedge Bel_iDes_j\varphi$ (from definition of $HappyFor'$)
2. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i(FBel_j\varphi \wedge GDes_j\varphi)$ ((by Pers-Des_i) and (C-MIX))
3. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_iF(Bel_j\varphi \wedge Des_j\varphi)$ (by property (1) for G)
4. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_iFJoy_j\varphi$ (by definition of Joy)

The proof is similar for cases (b) to (d). □

Theorem 4 (Consequences of fortunes-of-others emotions)

$$\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_iFBel_j\varphi \rightarrow Hope_iFBel_j\varphi \quad \text{(a)}$$

$$\vdash PSorryFor_{i,j}\varphi \wedge \neg Bel_iFBel_j\varphi \rightarrow Fear_iFBel_j\varphi \quad \text{(b)}$$

$$\vdash Resentment'_{i,j}\varphi \wedge \neg Bel_iFBel_j\varphi \rightarrow Fear_iFBel_j\varphi \quad \text{(c)}$$

$$\vdash Gloating'_{i,j}\varphi \wedge \neg Bel_iFBel_j\varphi \rightarrow Hope_iFBel_j\varphi \quad \text{(d)}$$

Proof 5 (of Theorem 4). Case of (a).

1. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i FBel_j\varphi \wedge Des_i Bel_j\varphi$ (from definition of $HappyFor'$)
2. $\vdash HappyFor'_{i,j}\varphi \rightarrow Prob_i FBel_j\varphi \wedge Des_i FBel_j\varphi$ (by contraposition of (T-G), and (RM-$\Box$) for Des_i)
3. $\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi \rightarrow Prob_i FBel_j\varphi \wedge \neg Bel_i FBel_j\varphi \wedge Des_i FBel_j\varphi$ (by $\mathcal{PL}$)
4. $\vdash HappyFor'_{i,j}\varphi \wedge \neg Bel_i FBel_j\varphi \rightarrow Expect_i FBel_j\varphi \wedge Des_i FBel_j\varphi$ (by definition 1)
5. $\vdash HappyFor'ij\varphi \wedge \neg Bel_i FBel_j\varphi \rightarrow Hopei FBel_j\varphi$ (by definition of $Hope$)

The proof is similar for cases (b) to (d). □

Theorem 5 (Other-agent emotions towards oneself)

$$\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Pride_i(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\vdash Reproach_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Shame_i(i{:}\alpha, \varphi) \quad \text{(b)}$$

Proof 6 (of Theorem 5). Case of (a). The proof comes immediately from the definitions of these two emotions.

1. $\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Bel_i\ Done_{i:\alpha}(\neg Prob_i\ Happens_{i:\alpha}\top \wedge Bel_i Idl_i\ Happens_{i:\alpha}\top)$ (by definition of $Admiration$)
2. $\vdash Admiration_{i,i}(i{:}\alpha, \varphi) \leftrightarrow Pride_i(i{:}\alpha, \varphi)$ (by definition of $Pride$)

The proof is similar for (b). □

Theorem 6 (Other-agent emotion does not force self-agent emotion)

$$\nvdash Bel_i Admiration_{j,i}(i{:}\alpha, \varphi) \rightarrow Pride_i(i{:}\alpha, \varphi) \quad \text{(a)}$$

$$\nvdash Bel_i Reproach_{j,i}(i{:}\alpha, \varphi) \rightarrow Shame_i(i{:}\alpha, \varphi) \quad \text{(b)}$$

Sketch of proof 1 (of Theorem 6) *It suffices to find a counter-example, i.e. a model where the implication is not valid, i.e. a model containing at least one world where the implication is false.*

Case of (b). By definition, $Bel_j Reproach_{i,j}(j{:}\alpha, \varphi)$ does not imply $Des_j \neg Happens_{j:\alpha}\varphi$. In a world where the first formula is true and the second one is false, the implication is false. For example, a teacher in a school can reproach to a student to wear unauthorised clothes, and tell this to him, without making this student ashamed of wearing them.

Theorem 7 (Link between prospect and attribution emotions)

$$\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}((\neg Bel_i \neg Happens_{i:\alpha}\varphi \wedge Des_i\ Happens_{i:\alpha}\varphi) \rightarrow Fear_i \neg Happens_{i:\alpha}\varphi) \quad \text{(a)}$$

$$\vdash Shame_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}((\neg Bel_i \neg Happens_{i:\alpha}\varphi \wedge Des_i \neg Happens_{i:\alpha}\varphi) \rightarrow Hope_i \neg Happens_{i:\alpha}\varphi) \quad \text{(b)}$$

$$\vdash Admiration_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{j:\alpha}((\neg Bel_i \neg Happens_{j:\alpha}\varphi \wedge Des_i\ Happens_{j:\alpha}\varphi) \rightarrow Fear_i \neg Happens_{j:\alpha}\varphi) \quad \text{(c)}$$

$$\vdash Reproach_{i,j}(j{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{j:\alpha}((\neg Bel_i \neg Happens_{j:\alpha}\varphi \wedge Des_i \neg Happens_{j:\alpha}\varphi) \rightarrow Hope_i \neg Happens_{j:\alpha}\varphi) \quad \text{(d)}$$

Proof 7 (of Theorem 7). Case of (a).

1. $\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}(Prob_i\ After_{i:\alpha} \neg\varphi)$ (by definition of $Pride$)
2. $\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}(Prob_i \neg Happens_{i:\alpha}\varphi)$ (by definition of $Happens$)
3. $\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i Done_{i:\alpha}(\neg Bel_i \neg Happens_{i:\alpha}\varphi \rightarrow Expect_i \neg Happens_{i:\alpha}\varphi)$ (by $\mathcal{PL}$ and definition 1)
4. $\vdash Pride_i(i{:}\alpha, \varphi) \rightarrow Bel_i\ Done_{i:\alpha}(\neg Bel_i \neg Happens_{i:\alpha}\varphi \wedge Des_i\ Happens_{i:\alpha}\varphi \rightarrow Fear_i \neg Happens_{i:\alpha}\varphi)$ (by $\mathcal{PL}$ and definition of $Fear$)

The proof is similar for (b), (c) and (d). □

Theorem 8 [Link between attribution and prospect emotions]. *If α is an action that the agent i believes to influence the proposition φ (cf. Remark 2), then:*

$$\vdash Fear_i \neg Happens_{i:\alpha}\varphi \wedge Bel_i Idl_i Happens_{i:\alpha}\varphi \rightarrow After_{i:\alpha}(Bel_i\varphi \rightarrow Pride_i(i{:}\alpha, \varphi)) \quad \text{(a)}$$

$$\vdash Hope_i \neg Happens_{i:\alpha}\varphi \wedge Bel_i Idl_i \neg Happens_{i:\alpha}\varphi \rightarrow After_{i:\alpha}(Bel_i\varphi \rightarrow Shame_i(i{:}\alpha, \varphi)) \quad \text{(b)}$$

$$\vdash Fear_i \neg Happens_{j:\alpha}\varphi \wedge Bel_i Idl_j \neg Happens_{j:\alpha}\varphi \rightarrow After_{j:\alpha}(Bel_i\varphi \rightarrow Admiration_{i,j}(j{:}\alpha, \varphi)) \quad \text{(c)}$$

$$\vdash Hope_i \neg Happens_{j:\alpha}\varphi \wedge Bel_i Idl_j \neg Happens_{j:\alpha}\varphi \rightarrow After_{j:\alpha}(Bel_i\varphi \rightarrow Reproach_{i,j}(j{:}\alpha, \varphi)) \quad \text{(d)}$$

To prove Theorem 8 we need the following lemma.

Lemma 2 $Done_\alpha \neg Bel_i\ After_\alpha \bot \wedge Done_\alpha Bel_i\varphi \rightarrow Bel_i\ Done_\alpha\varphi$

To prove Lemma 2 we need the following lemma.

Lemma 3 *if* $\varphi \rightarrow After_\alpha\psi$ *then* $Done_\alpha\varphi \rightarrow \psi$

Proof 8 (of Lemma 3).

1. $\varphi \rightarrow After_\alpha\psi$ (by hypothesis)

2. $Done_\alpha\, After_\alpha \varphi \rightarrow \varphi$ (from contraposition of (CONV-BH))
3. $Done_\alpha \varphi \rightarrow Done_\alpha\, After_\alpha \psi$ (from 1. by (RM-$\Diamond$) for $Done_\alpha$)
4. $Done_\alpha \varphi \rightarrow \psi$ (from 2. and 3.)□

Proof 9 (of Lemma 2). 1. $Bel_i\, After_\alpha \varphi \wedge \neg Bel_i After_\alpha \bot \rightarrow After_\alpha Bel_i \varphi$ (from (NF-Bel_i))
2. $Bel_i\, After_\alpha\, Done_\alpha \varphi \wedge \neg Bel_i After_\alpha \bot \rightarrow After_\alpha Bel_i Done_\alpha \varphi$ (by instantiation of 1.)
3. $\varphi \rightarrow After_\alpha\, Done_\alpha \varphi$ (from (CONV-AD))
4. $Bel_i \varphi \rightarrow Bel_i\, After_\alpha\, Done_\alpha \varphi$ (from 3. by (RM-□) for Bel_i)
5. $Bel_i \varphi \wedge \neg Bel_i\, After_\alpha \top \rightarrow After_\alpha Bel_i\, Done_\alpha \varphi$ (from 2. and 4. by $\mathcal{PL}$)
6. $Done_\alpha (Bel_i \varphi \wedge \neg Bel_i\, After_\alpha \top) \rightarrow Bel_i\, Done_\alpha \varphi$ (by Lemma 3)
7. $Done_\alpha \Phi \wedge Done_\alpha \Psi \rightarrow Done_\alpha (\Phi \wedge \Psi)$ (from (CD-DB))
8. $Done_\alpha \neg Bel_i\, After_\alpha \bot \wedge Done_\alpha Bel_i \varphi \rightarrow Bel_i\, Done_\alpha \varphi$ (from 6. and 7.) □

Proof 10 (of Theorem 8). Case of (a).

1. $Fear_i \neg Happens_{i:\alpha} \varphi \rightarrow \neg Bel_i \neg Happens_{i:\alpha} \varphi$ (by definition of $Fear$ and definition 1)
2. $Fear_i \neg Happens_{i:\alpha} \varphi \rightarrow \neg Bel_i \neg Happens_{i:\alpha} \top$ (from 1. by (RM-$\Diamond$) for $\neg Bel_i \neg$)
3. $Fear_i \neg Happens_{i:\alpha} \varphi \rightarrow \neg Bel_i After_\alpha \bot$ (from 2. by definition of $Happens$)
4. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow Bel_i (Fear_i \neg Happens_{i:\alpha} \varphi \wedge Idl_i\, Happens_{i:\alpha} \varphi \wedge \neg Bel_i After_\alpha \bot)$ (by Theorem 13, (5-Bel_i) and (C-□) for Bel_i)
5. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow After_{i:\alpha}\, Done_{i:\alpha} Bel_i (Fear_i \neg Happens_{i:\alpha} \varphi \wedge Idl_i\, Happens_{i:\alpha} \varphi \neg Bel_i\, After_\alpha \bot)$ (by (CONV-AD))
6. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow After_{i:\alpha} Bel_i\, Done_{i:\alpha} (Fear_i \neg Happens_{i:\alpha} \varphi \wedge Idl_i\, Happens_{i:\alpha} \varphi)$ (by Lemma 2)
7. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow After_{i:\alpha} (Bel_i \varphi \rightarrow Bel_i\, Done_{i:\alpha} (Fear_i \neg Happens_{i:\alpha} \varphi \wedge Idl_i\, Happens_{i:\alpha} \varphi) \wedge Bel_i \varphi)$
8. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow After_{i:\alpha} (Bel_i \varphi \rightarrow Bel_i\, Done_{i:\alpha} (Prob_i\, After_{i:\alpha} \neg \varphi \wedge Idl_i\, Happens_{i:\alpha} \varphi) \wedge Bel_i \varphi)$ (by definitions of $Fear$ and $Happens$)
9. $\vdash Fear_i \neg Happens_{i:\alpha} \varphi \wedge Bel_i Idl_i\, Happens_{i:\alpha} \varphi \rightarrow After_{i:\alpha} (Bel_i \varphi \rightarrow Pride_i (i{:}\alpha, \varphi))$ (by definition of $Pride$)

The proof is similar for (b), (c), and (d). □

Theorem 9 (Inconsistency between good-will and ill-will emotions)

$$\vdash \neg (HappyFor_{i,j} \varphi \wedge Resentment_{i,j} \varphi) \quad \text{(a)}$$

$$\vdash \neg (SorryFor_{i,j} \varphi \wedge Gloating_{i,j} \varphi) \quad \text{(b)}$$

$$\vdash \neg (HappyFor'_{i,j} \varphi \wedge Gloating_{i,j} \varphi) \quad \text{(c)}$$

$$\vdash \neg (SorryFor_{i,j} \varphi \wedge Resentment_{i,j} \varphi) \quad \text{(d)}$$

Sketch of proof 2 (of Theorem 9) *The proof for cases (a) and (b) follows from the rationality of* Des_i. *The proof for cases (c) and (d) follows from Lemma 4.* □

Lemma 4 $\neg(Bel_i\, Des_j \varphi \wedge Bel_i\, Des_j \neg\varphi)$

Proof 11 (of Lemma 4).

1.	$\vdash Des_j\varphi \rightarrow \neg Des_j\neg\varphi$	(from (D-Des_i))
2.	$\vdash Bel_i\, Des_j\varphi \rightarrow Bel_i\neg Des_j\neg\varphi$	(by (RM-$\Box$) for Bel_i)
3.	$\vdash Bel_i\, Des_j\varphi \rightarrow \neg Bel_i\, Des_j\neg\varphi$	(by (D-Bel_i))
4.	$\vdash \neg(Bel_i\, Des_j\varphi \wedge Bel_i\, Des_j\neg\varphi)$	(by $\mathcal{PL}$)

□

References

Adam, C. (2007). *Emotions: From psychological theories to logical formalization and implementation in a BDI agent*. Ph.D. thesis, INP Toulouse, France (available in English).

Adam, C., Gaudou, B., Herzig, A., & Longin, D. (2006). OCC's emotions: A formalization in a BDI logic. In J. Euzenat (Ed.), *Proceedings of the twelfth internatinal conference on artificial intelligence: Methodology, systems, and applications (AIMSA'06), Varna, Bulgaria, september 13–15*, Vol. 4183 of *LNAI* (pp. 24–32). Springer-Verlag.

Adam, C., Gaudou, B., Longin, D., & Lorini, E. (2009). Logical modeling of emotions for Ambient Intelligence. Technical report of the institution IRIT, Toulouse. Available online at http://www.irit.fr/-Publications.

Adam, C., & Longin, D. (2007). Endowing emotional agents with coping strategies: From emotions to emotional behaviour. In Catherine Pelachaud et al. (Eds.): *Intelligent Virtual Agents (IVA'07)*, Vol. 4722 of *LNCS* (pp. 348–349). Springer-Verlag.

Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *Proceedings of the fourth IEEE international conference on multimodal interfaces (ICMI 2002)* (pp. 483–490). IEEE Computer Society.

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Arnold, M. (1960). *Emotion and personality*. New York: Columbia University Press.

Åqvist, L. (2002). Deontic logic. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic*, (Vol. 8, 2nd ed., pp. 147–264). Kluwer Academic Publishers.

Bartneck, C. (2002). *eMuu—an embodied emotional character for the ambient intelligent home*. Ph.D. thesis, Eindhoven University of Technology.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM, 37*(7), 122–125.

Becker, C., Kopp, S., & Wachsmuth, I. (2004). Simulating the emotion dynamics of a multimodal conversational agent. In *ADS'04*. Springer LNCS.

Bower, G. H. (1992). How might emotions affect learning. In S.-A. Cristianson (Ed.), *The handbook of emotion and memory: Research and theory* (pp. 3–32). Lawrence Erlbaum Associates.

Bratman, M. E. (1987). *Intention, plans, and practical reason*. Cambridge, MA, USA: Harvard University Press.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*, 161–178.

Burgess, J. P. (1969). Probability logic. *Journal of Symbolic Logic, 34*, 264–274.

Burgess, J. P. (2002). Basic tense logic. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (Vol. 7, 2nd ed., pp. 1–42). Kluwer Academic Publishers.

Castelfranchi, C., & Paglieri, F. (2007). The role of belief in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese, 155*, 237–263.

Chellas, B. F. (1980). *Modal logic: An introduction*. Cambridge: Cambridge University Press.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence Journal, 42*(2–3), 213–261.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam Pub Group.

Darwin, C. R. (1872). *The expression of emotions in man and animals*. London: Murray.
Dastani, M., & Meyer, J. -J. (2006). Programming agents with emotions. In *Proceedings 17th European conference on artificial intelligence (ECAI 2006), Trento, Italy, Aug. 28th–Sep. 1st*. IOS Press.
de Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., & Carolis, B. D. (2003). From Greta's mind to her face: Modelling the dynamics of affective states in a conversational embodied agent. *International Journal of Human-Computer Studies, 59*(1–2), 81–118.
Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200.
Ekman, P., Friesen, W., & Hager, J. (2002). *Facial action coding system invetigator's guide*. A Human Face.
Elliott, C. (1992). *The affective reasoner : A process model of emotions in a multi-agent system*. Ph.D. thesis, Northwestern University, Illinois.
Elliott, C., Rickel, J., & Lester, J. (1999). Lifelike pedagogical agents and affective computing: An exploratory synthesis. *Lecture Notes in Computer Science, 1600*, 195–211.
Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature, 36*(1), 47–74.
Fagin, R., & Halpern, J. Y. (1994). Reasoning about knowledge and probability. *Journal of the ACM* **41**(2), 340–367.
Fagin, R., Halpern, J. Y., Vardi, M. Y., & Moses, Y. (1995). *Reasoning about knowledge*. Cambridge: MIT Press.
Forgas, J. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin, 117*, 39–66.
Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
Gershenson, C. (1999). Modelling emotions with multidimensional logic. In *NAFIPS'99*. IEEE.
Gordon, R. (1987). *The structure of emotions*. New York: Cambridge University Press.
Gratch, J., & Marsella, S. (2004). A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research, 5*(4), 269–306.
Grizard, A., & Lisetti, C. (2006). Generation of facial emotional expressions based on psychological theory. In *Workshop on Emotion and Computing, KI'2006*. Bremen, Germany.
Halpern, J., & McAllester, D. (1989). Likelihood, probability, and knowledge. *Computational Intelligence, 5*, 151–160.
Halpern, J., & Rabin, M. (1987). A logic to reason about likelihood. *Artificial Intelligence Journal, 32*(3), 379–405.
Herzig, A. (2003). Modal probability, belief, and actions. *Fundamenta Informaticæ, 57*(2–4), 323–344.
Herzig, A., & Longin, D. (2002). Sensing and revision in a modal logic of belief and action. In F. van Harmelen (Ed.), *Proceedings of 15th European conference on artificial intelligence (ECAI 2002), Lyon, France, July 23–26* (pp. 307–311). IOS Press.
Herzig, A., & Longin, D. (2004). C&L intention revisited. In D. Dubois, C. Welty, & M. -A. Williams (Eds.), *Proceedings of the 9th international conference on principles of knowledge representation and reasoning (KR 2004), Whistler, Canada, June 2–5* (pp. 527–535). AAAI Press.
Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.
Jaques, P. A., Vicari, R. M., Pesty, S., & Bonneville, J. -F. (2004). Applying affective tactics for a better learning. In *Proceedings of the 16th European conference on artificial intelligence (ECAI 2004)*. IOS Press.
Klein, J., Moon, Y., & Picard, R. (1999). This computer responds to user frustration. In *Proceedings of the conference on human factors in computing systems* (pp. 242–243). Pittsburgh, USA: ACM Press.
Lang, J., van Der Torre, L. W. N., & Weydert, E. (2002). Utilitarian desires. *Journal of Autonomous Agents and Multi-Agent Systems, 5*, 329–363.
Laverny, N., & Lang, J. (2005). From knowledge-based programs to graded belief-based programs, Part II: Off-line reasoning. In *Proceedings of the 9th international joint conference on artificial intelligence (IJCAI'05) , Edinburgh, Scotland, 31/07/05-05/08/05* (pp. 497–502). Gallus.
Lazarus, R. (1999). The cognition–emotion debate: A bit of history. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 3–20). New York: Wiley.
Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford University Press.
Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer Publishing Company.
Lenzen, W. (1978). *Recent work in epistemic logic*. Amsterdam: North Holland Publishing Company.
Lenzen, W. (1995). On the semantics and pragmatics of epistemic attitudes. In A. Laux & H. Wansing (Eds.), *Knowledge and belief in philosophy and AI* (pp. 181–197). Berlin: Akademie Verlag.

Loewenstein, G. (2000). Emotions in economic theory and economic behavior. *American Economic Review, 90*(2), 426–432.
Lorini, E., & Castelfranchi, C. (2006). The unexpected aspects of Surprise. *International Journal of Pattern Recognition and Artificial Intelligence, 20*(6), 817–833.
Lorini, E., & Castelfranchi, C. (2007). The cognitive structure of Surprise: looking for basic principles. *Topoi: An International Review of Philosophy, 26*(1), 133–149.
Lorini, E., & Herzig, A. (2008). A logic of intention and attempt. *Synthese, 163*(1), 45–77.
Meyer, J. J. (2004). Reasoning about emotional agents. In R. L. de Mántaras & L. Saitta (Eds.), *16th European conference on artificial intelligence (ECAI)* (pp. 129–133).
Meyer, J. -J. (2006). Reasoning about emotional agents. *International Journal of Intelligent Systems, 21*(6), 601–619.
Meyer, J. -J., de Boer, F., van Eijk, R., Hindriks, K., & van der Hoek, W. (2001). On programming Karo agents. *Logic Journal of the IGPL, 9*(22), 261–272.
Meyer, J. J. C., van der Hoek, W., & van Linder, B. (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence, 113*(1–2), 1–40.
Meyer, W. U., Reisenzein, R., & Schützwohl, A. (1997). Towards a process analysis of emotions: The case of surprise. *Motivation and Emotion, 21*, 251–274.
Ochs, M., Niewiadomski, R., Pelachaud, C., & Sadek, D. (2005). Intelligent expressions of emotions. In *1st international conference on affective computing and intelligent interaction ACII*. China.
Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge, MA: Cambridge University Press.
Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with computers, 16*, 295–309.
Parunak, H., Bisson, R., Brueckner, S., Matthews, R., & Sauter, J. (2006). A model of emotions for situated agents. In P. Stone & G. Weiss (Eds.), *AAMAS'06* (pp. 993–995). ACM Press.
Picard, R., & Liu, K. (2007). Relative subjective count and assessment of interruptive technologies applied to mobile monitoring of stress. *International Journal of Human-Computer Studies, 65*, 396–375.
Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence, 19*, 297–285.
Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In J. A. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings second international conference on principles of knowledge representation and reasoning (KR'91)* (pp. 473–484). Morgan Kaufmann Publishers.
Rao, A. S., & Georgeff, M. P. (1992). An abstract architecture for rational agents. In B. Nebel, C. Rich, & W. Swartout (Eds.), *Proceedings third international conference on principles of knowledge representation and reasoning (KR'92)* (pp. 439–449). Morgan Kaufmann Publishers.
Reilly, N. (1996). *Believable social and emotional agents*. Ph.D. thesis. Pittsburgh, PA, USA.: School of Computer Science, Carnegie Mellon University.
Reiter, R. (1991). The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz (Ed.), *Artificial intelligence and mathematical theory of computation: Papers in honor of John McCarthy* (pp. 359–380). San Diego, CA: Academic Press.
Russell, J. A. (1997). How shall an emotion be called? In R. Plutchik & H. Conte (Eds.), *Circumplex models of personality and emotions* (pp. 205–220). Washington, DC: American Psychological Association.
Sadek, M. (1992). A study in the logic of intention. In B. Nebel, C. Rich, & W. Swartout (Eds.), *Proceedings third international conference on principles of knowledge representation and reasoning (KR'92)* (pp. 462–473). Morgan Kaufmann Publishers.
Scherer, K. (1987). Toward a dynamic theory of emotion: the component process model of affective states. *Geneva studies in Emotion and Communication, 1*(1), 1–98.
Scherer, K. R. (2001). *Appraisal processes in emotion: Theory, methods, research*, Chapt. Appraisal considered as a process of multilevel sequential checking (pp. 92–120). New York: Oxford University Press.
Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. New York: Cambridge University Press.
Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
Segerberg, K. (1971). Qualitative probability in a modal setting. In J. Fenstad (Ed.), *Proceedings of the 2nd Scandinavian logic symposium*. Amsterdam, North Holland Publishing Company.
Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge. *Journal of personality and Social Psychology, 52*, 1061–1086.

Solomon, R., & Calhoun, C. (Eds.), (1984). *What is an emotion? Classic readings in philosophical psychology*. Oxford: Oxford University Press.

Staller, A., & Petta, P. (2001). Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation, 4*(1). Available online at http://jasss.soc.surrey.ac.uk/4/1/2.html.

Steunebrink, B. R., Dastani, M., & Meyer, J. J. C. (2007). A logic of emotions for intelligent agents. In *Proceedings of the twenty-second AAAI conference on artificial intelligence (AAAI'07)* (pp. 142–147). AAAI Press.

van der Hoek, W., van Linder, B., & Meyer, J.-J. C. (1997). An integrated modal approach to rational agents. In *Proceedings of 2nd AISB workshop on practical reasoning and rationality* (pp. 123–159). Manchester, United Kingdom.

Walley, P., & Fine, T. L. (1979). Varieties of modal (classificatory) and comparative probability. *Synthese, 41*, 321–374.