Emotions

Carole Adam

THÈSE

présentée devant

L' INSTITUT NATIONAL POLYTECHNIQUE DE TOULOUSE

pour obtenir le grade de

Docteur de l'Institut National Polytechnique de Toulouse
Discipline : Intelligence Artificielle

présentée et soutenue par

Carole Adam

le : mardi 10 Juillet 2007

Titre :

Emotions: from psychological theories to logical formalization and implementation in a BDI agent

---

Directeur de thèse :
Andreas Herzig

Encadrants :
Dominique Longin
Fabrice Evrard

Laboratoire d'accueil :
Institut de Recherche en Informatique de Toulouse

JURY

Yves Demazeau[2,3]	CR CNRS, Laboratoire d'Informatique de Grenoble
Robert Demolombe	DR associé CNRS, IRIT, Toulouse
Fabrice Evrard	Maître de Conférences, ENSEEIHT, Toulouse
Andreas Herzig	DR CNRS, IRIT, Toulouse
Dominique Longin	CR CNRS, IRIT, Toulouse
Emiliano Lorini	Docteur-ingénieur, IRIT Toulouse - ISTC Rome, Italie
John-Jules Meyer[2]	Professeur, Université Utrecht, Pays-Bas
Catherine Pelachaud[1,2]	Professeur, IUT Montreuil, Université Paris 8
David Sadek	Docteur-ingénieur, France Telecom R&D, Lannion

[1] Présidente du jury
[2] Rapporteur
[3] Absent le jour de la soutenance

Remerciements

Pour commencer cette thèse, voici la traditionnelle page de remerciements, qui a nécessité plus de temps d'écriture que tout le reste du manuscrit...

Tout d'abord je voudrais remercier mes rapporteurs, Yves Demazeau, John-Jules Meyer, et Catherine Pelachaud pour avoir accepté de relire mon manuscrit et pour leurs nombreux commentaires qui m'ont (j'espère) aidée à l'améliorer. Merci à eux ainsi qu'aux autres membres du jury, David Sadek, Emiliano Lorini, et Robert Demolombe, pour avoir assisté à ma soutenance (ou avoir essayé) et m'avoir posé nombre de questions intéressantes.

Je voudrais ensuite remercier mes encadrants, Andreas Herzig, Dominique Longin, et Fabrice Evrard, pour m'avoir encadrée et supportée pendant quatre ans, pour m'avoir motivée ou freinée selon les jours, pour leurs idées, leurs critiques et leurs commentaires.

Merci aussi aux personnes avec qui j'ai eu l'occasion de travailler et d'échanger pendant cette thèse : l'équipe GRIC de l'IRIT (Bernard Pavard, Nico Pallamin, Mehdi El Jed, Lucila Morales), Robert Demolombe, Nicolas Asher...

Merci au Groupe de Travail sur les Agents Conversationnels Animés qui m'a donné la première occasion d'exposer mes travaux en dehors de l'IRIT, et m'a permis de rencontrer et d'échanger avec d'autres chercheurs intéressés par les mêmes sujets. Merci en particulier à Sylvie Pesty et Jean-Paul Sansonnet pour leurs conseils et leur soutien.

Merci à tous ceux qui m'ont fait confiance pour organiser WACA'2006, en particulier les responsables du comité scientifique Catherine Pelachaud et Jean-Claude Martin. Merci aussi à ceux sans l'aide de qui je n'aurais pas pu l'organiser : Véronique Desbats et Dominique Longin. L'organisation de ce workshop a été une très bonne expérience, même si la rédaction de ma thèse en fût quelque peu retardée...

Merci à l'équipe enseignante de l'ENSEEIHT de m'avoir permis d'enseigner pendant trois années dont je garde un excellent souvenir. Merci à tous les élèves que j'ai vu passer pendant ces trois ans pour m'avoir donné goût à l'enseignement.

Merci à Domi, alias Dieu du Latex, pour avoir toujours su faire compiler des fichiers récalcitrants ; pour avoir toujours réussi à faire tenir un article de 12 pages dans les 10 pages imposées, sans forcément en raccourcir le texte ; pour m'avoir initiée à la typographie et aux joies de l'édition d'actes de conférence ; pour avoir assisté trois fois à la répétition de ma soutenance ; et pour tout le reste.

Merci à Andi pour avoir accepté de tenter la formalisation logique des émotions, pas forcément logique au premier abord ; pour avoir refusé d'y associer des intensités (et m'en avoir convaincue) ; pour sa disponibilité et ses connaissances insondables et indispensables ; pour les longues discussions passionnées dans son bureau ; et pour tout le reste.

Merci à ma soeur (et encore félicitations) de s'être mariée trois jours avant ma soutenance (le 7/7/7, quelle originalité...) pour m'empêcher de stresser. Merci aussi à ses invités d'avoir laissé suffisamment de bouteilles pour sustenter la soif de ceux de mon pot de thèse. Merci à mon autre soeur pour sa sauce au thon et sa tapenade aux olives. Merci à ma famille et belle-famille d'être venus assister à ma soutenance même sans rien y comprendre.

Merci à Galaad, mon fidèle ordinateur, pour ses valeureux services pendant quatre ans.

Merci à toutes les personnes que j'ai oublié de citer de faire preuve de tant de bienveillance à mon égard et de ne même pas m'en vouloir, ou alors pas longtemps...

Merci à l'hypothétique lecteur qui aurait pris le temps de parcourir cette page de remerciements.

Et enfin merci à Benoit pour m'avoir (plus ou moins) supportée pendant ces quatre ans de thèse, et pour tout le reste. Á mon tour maintenant ! (Alors cette rédaction, ça avance ? :-p)

Pour Odile

Contents

List of Figures

List of Tables

Introduction

Think about it: you have better access to your innermost feelings than anyone, but you still do not always know how to "recognize" or label what you are feeling.
(Picard, 2003, p. 2)

We humans feel emotions everyday, almost every minute, not always consciously. Our emotions have a central role in our lives, in our relations with others (sympathy, empathy, laugh, love...). They can be manipulated by advertisements to make us buy, by politicians who want us to vote for them, or by any people wanting to seduce us. They can make us perform impulsive actions that we regret later. They can make our life more difficult (*e.g.* shyness, stress...) and interfere with our work by distracting us. We often want to hide them, due to cultural norms (like in Japan where people express few emotions in public) or because they could reveal what we think (in business or when playing poker for example). Finally we usually believe that we have to control our emotions in order to be rational or to perform better. Incidentally athletes now have mental trainers who teach them how to control their emotions. Actually more and more people call on such mental coaches or read books supposed to help them control undesirable emotions.

Nevertheless controlling one's own emotions is all the more difficult as we do not even understand them. Actually we do not even know exactly what is an emotion. We would be incapable of giving a clear definition of this phenomenon. And yet, many people tried to find such a definition. But the difficulty is that emotions simultaneously involve several types of interdependent reactions: cognitive, physiological, biological... that relate to different fields of research. So most definitions of emotions are partial since they only refer to one of their aspects.

Emotions have long interested only philosophers, who often consider them as something man has to be expurgated of, for example through the concept of *catharsis* for Aristotle (2003). Spinoza (1994) associates emotions with inadequate ideas and passivity of mind, in opposition with reason. Descartes (1649) assimilates emotions (or passions) with animal instincts that human reason has to regulate.

Then biology assumed that emotions were useful reflexes inherited during evolution (Darwin, 1872), and proposed classifications of basic emotions. Physiology then tried to explain the emotional mechanisms and several theories emerged, disagreeing on the emotional locus that they identified (Lange and James, 1967; Cannon, 1927). Psychology agrees on the adaptive role of emotions in human life, and proposed several theories for explaining their triggering from a cognitive point of view (Schacter and Singer, 1962; Arnold, 1950). Sociology gets in turn interested in explaining emotions, and more particularly their social construction and their role in maintaining the cohesion of a group (Averill, 1980; Durkheim, 1961). But in spite of all these theories accounting for the essential role of emotions for individuals and for the society, computer science still neglects this phenomenon considered as complex and irrational, so irrelevant for rational agents.

Fortunately, the progress of neuroscience (Damasio, 1994) proves the role of emotions in intelligent behaviour, decision making, planning, social communication, and all these supposed rational human abilities: we humans cannot reason if we do not feel emotions. This tangible evidence begin to convince computer scientists of the usefulness of emotions for intelligent agents. Then Bates (1994) introduces the concept of *believable agents*, agents that can give the illusion of life, and shows that emotions are something crucial for such agents. Finally Picard (1997) is the first computer scientist to argue that virtual agents need emotions not only to be believable but also to be truly intelligent and to interact in a natural and friendly way with humans. She introduces the concept of Affective Computing, and computer science then starts being really interested in emotions. The agent community thus begins to design emotional agents for various applications: intuitive human-machine interfaces, believable agents for virtual worlds, intelligent tutoring agents... These agents must not only express emotions, but also perceive and understand those of the user to adapt their behaviour to them. Of course what is called emotions for these agents does not match exactly what we believe human emotions to be. The *virtual emotions* given to virtual agents are rather some kinds of labels concisely describing a particular state of this agent impacting his behaviour (Picard, 1997): when the agent is in this state he expresses a behaviour consistent with this emotion, even if he does not really "feel" it in our human sense.

The design of emotional agents must involve both computer science and psychology in cooperation (Gratch and Marsella, 2005). Indeed, psychologists have already tried to decode emotions for decades. Thus agents designers try to find a comprehensible psychological theory that is adapted to their objective. Most of them build on Ortony, Clore and Collins' typology of emotions (Ortony, Clore,

and Collins, 1988) that was intended to be used in AI applications. They then provide their own formalization or implementation of (part of) this theory in order to integrate it in their agents.

Now we believe that understanding and formalizing the theory is a hard work that should be done once and for all. It would be a waste of time and energy to start from scratch each time someone needs to implement an emotional agent. Moreover, direct implementations of a theory are not reusable for other agents or applications. Thus it is important to propose generic reusable models ready to be implemented in agents, and we assume that formal logics offer the required properties to design such models.

Indeed we just showed that emotions are a complex phenomena whose definition is often abstract, ambiguous, subjective and non consensual. On the contrary, emotional agent designers need clear definitions, ready to be implemented in their agents. Formal logics provide such a universal vocabulary, with a clear semantics. They also allow reasoning, planning and explanation of an agent's behaviour. The logical formalization of a phenomenon can even reveal problems that do not appear intuitively. BDI logics, *viz.* logics of *Belief*, *Desire* and *Intention* (Cohen and Levesque (1990), Rao and Georgeff (1991, 1992), Sadek (1992), Herzig and Longin (2004), Wooldridge (2000)) ground on the philosophy of language, mind, and action (*cf.* Bratman (1987), Searle (1969, 1983)). They propose to model agents *via* some key concepts such as action and *mental attitudes* (beliefs, goals, intentions, choices...). This framework is commonly used in the agent community and offers well-known interesting properties: great explanatory power, formal verifiability, rigorous and well-established theoretical frame (from the point of view of both philosophy and formal logic).

Besides, Searle (1983) assumes that emotions are particular mental attitudes and can be expressed in terms of beliefs and desires, what supports the idea of using BDI logics to represent them. Nevertheless there is not much work about emotions in this area yet. For example Meyer (2004) proposes a language for the description of emotional agents based on a BDI logic, and Ochs et al. (2005) also provide definitions of some emotions using Sadek's rational interaction theory (1992), a particular BDI formalism. But these models do not take advantage of all the assets of BDI logics. In particular they are more interested in the implementation of their model in an agent than in the formal reasoning possibilities that it allows. Besides it seems to be fair to say that they are not faithful enough to the psychological theories and do not propose a rich enough set of emotions.

Therefore our objective in this thesis is to provide a generic logical model of emotions. Actually we are only interested in the cognitive aspect of emotions, at the expense of the biological and physiological aspects. We do not aim at provid-

ing an agent architecture or implementation, but rather at disambiguating emotions and reason about their properties. Thus our work is not at the same level than most of existing computer science approaches to emotions, but rather proposes to be the basis of their implementations. There already exist some formalizations of emotions, that ground on various formalisms. We will thus argue our choice to use BDI logics rather than other existing formalisms to design our model. We will then discuss two existing formalisms that also ground on BDI logics, insisting on what we propose to improve in these approaches. In particular our model should be faithful to a psychological theory, provide a richer set of emotions than existing models, and have a correct and complete axiomatics to allow deductions about emotions. Indeed, we want to use our model to prove some theorems about emotions, what is very interesting and was never done before, as far as we know. However this model should not be restricted to such a theoretical use, what is often criticized in formal logics, so we also want to use it in practical applications and to implement it in a BDI agent. This model should provide believable emotions to these agents so we will finally assess this point.

To summarize, the main objective of our logical model, and its main advantage over existing formalizations, is to disambiguate emotions as exhaustively as possible, by representing a great number of them in a formal language, and to allow reasoning about their properties.

This thesis is structured in three parts. The first part is dedicated to the state of the art, from the point of view of both psychology and computer science. In Chapter 1 we answer the essential question of "what is an emotion?" by introducing some psychological theories of emotions in a historical perspective. Confronted to the great variety of existing theories, a second question arises: "which theory is best adapted to be formalized?", that we will also answer there. This theory will be described in more details than the other ones. Then, to convince readers wondering whether emotions are really useful for virtual agents, we will expose some key research findings that initiated the interest of computer science in emotions, and describe some existing Artificial Intelligence applications where emotions improve the efficacy of intelligent agents in their task (Chapter 2).

The second part sets up the core of this thesis, *viz*. our formalization of emotions. We introduce our particular BDI formalism, give its semantics and axiomatics, and prove its soundness and completeness (Chapter 3). We then formalize twenty emotions in terms of the modal operators thus described; we discuss our choices of formalization as well as the differences between several psychological theories describing these emotions or other close emotions (Chapter 4). Once our definitions are accepted (and we hope that our arguments can help this), our logic allows us to prove some intuitive properties of emotions (Chapter 5). This supports

the accuracy of our definitions, as well as the power of BDI logics to disambiguate complex concepts.

Finally the third part exposes some concrete applications and ongoing work following from this formalization. In particular we show some small case studies on Ambient Intelligence and machine-machine dialogue (Chapter 6). We then describe an implementation of our formalism in the agent PLEIAD and its use for assessing the believability of the generated emotions (Chapter 7). Finally we expose preliminary results about the formalisation of *coping*, the process describing how an agent can manage his negative emotions, in the same logical formalism (Chapter 8).

Part I

Emotions: from psychology to computer science

The beginning of knowledge is the discovery of something we do not understand

Frank Herbert

Chapter 1

Emotions in psychology

The great tragedy of science – the slaying
of a beautiful hypothesis by an ugly fact
Thomas Huxley, biologist and writer.

1.1 Short history

Emotions have always been subject to great debate. They were first only investigated by philosophy and assumed to be disorganizing and irrational. Darwin designed the first modern theory of emotions, recognizing their essential role in survival and their adaptive function. Several researchers designed other evolutionary theories, proposing different sets of basic emotions with various including criteria. This variety in the number and names of the basic emotions considered led to criticism and doubt on the very existence of basic emotions (Ortony and Turner, 1990).

Besides, the first cognitive theories appear. Arnold (1960) introduces the concept of cognitive appraisal as prior to any emotion: an emotion can only be triggered by the awareness of the reactions ordered by the brain in response to the appraisal of the situation. Almost at the same moment Schacter and Singer (1962) also propose to integrate cognition in the emotional triggering process: they suggest that an emotion arises from the analysis of physiological signals, but these signals must be cognitively disambiguated first. The concept of appraisal is agreed on by Lazarus (1966) who says that stress depends on the meaning of the stimulus for the individual perceiving it. This importance given to cognition makes emotions too deliberative and was much criticized since. Zajonc (1980) exposes experiments proving that emotions can arise without any cognitive activity. Lazarus (1984a) answers by showing that the same situation can trigger different emotions depending

on how the individual appraises it. A great debate then opposed the sustainers of the primacy of cognition (Lazarus, 1984b) and those of the primacy of affect (Zajonc, 1984) during the 80s. Leventhal and Scherer (1987) then propose to distinguish between three levels of cognitive appraisal: the sensorimotor level is innate and unconscious; the schematic level is acquired but still unconscious; finally only the conceptual level is conscious. Provided this distinction, we can admit that emotions really involve cognitions but of more or less high level. The debate may then have arisen from a disagreement on the definition of cognition. In the sequel cognitive appraisal theories become an active branch of psychology, and the concept of appraisal is developed by numerous researchers (Frijda, 1986; Scherer, 1987; Ortony, Clore, and Collins, 1988; Lazarus, 1991).

In this chapter we give an overview of the main psychological trends in the theorization of emotions. We then highlight the assets of one particular type of psychological theories of emotions: the cognitive appraisal theories. We give more details on two theories of this research trend that particularly interest us in this thesis: Ortony, Clore and Collins' typology, and Lazarus' theory. Finally we will conclude about the emerging interest of computer science in the psychology of emotions.

1.2 Main trends in psychological research on emotions

As the history showed, several trends of research oppose to each other to answer some big questions about emotions. Actually two questions mainly interest psychology of emotions:

1. **classification**: can we find some basic emotions or emotional categories? Different types of approaches consider (for various reasons) that there effectively exists a limited set of basic emotions, while continuous theories are opposed to this view (*cf.* Section 1.2.1);

2. **functioning**: how are emotions triggered? What is their influence on behaviour? Does cognition play a role? Do physiological changes trigger emotions or do emotions trigger physiological changes? Physiological (*cf.* Section 1.2.2) and cognitive (*cf.* Section 1.2.3) theories take opposite views on these questions.

Below we list what we believe to be the main trends in psychological research on emotion, and give a few details about some theories in each trend. We also summarize in tables the answers that each theory proposes to these essential questions, in order to make it easier to compare these theories. Actually these theories

sometimes answer only one of these questions and take an implicit position or no position at all on the other one. Table 1.1 summarizes these trends of research and their answers to the big questions presented above ("-" means that this theory does not explicitly answer the question or is not concerned with it).

Table 1.1: Different trends in the psychology of emotions

	Basic emotions?	**Cognitive component?**
Discrete theories	Yes	-
• Evolutionary theories	Yes, adaptive	No
• Non-evolutionary	Yes, biologic	No
• Building blocks	Yes, primary	-
Continuous theories	No	-
Physiological theories	-	No
Cognitive theories	-	Yes
• Two-factor	-	Interpretation of arousal
• Appraisal	-	Appraisal of stimuli

1.2.1 Theories interested in the representation of emotions: discrete *vs* continuous

The first important question that psychology tried to answer about emotions was their classification. Two different trends take opposite views on this problem: the first one assumes that there exists a limited set of basic emotions, while the second one considers emotions as a continuous function of two or three dimensions. In the "discrete trend", Ekman (1999a) distinguishes three different definitions of basic emotions:

- adaptive emotions in relation with an evolution theory: the basic emotions are those designed along evolution to solve specific problems, and are now innate answers programmed in individuals (*e.g.* Darwin, Ekman, Oatley and Johnson-Laird);

- discrete emotions that importantly differ one from another (not inevitably in relation with an evolution theory, *e.g.* Tomkins, Izard); this view is opposed to continuous approaches considering that emotions are all similar in essence and only differ by the values of some dimensions like intensity or valence (*e.g.* Lang, Russell);

- building blocks: this is Plutchik's view, comparing basic emotions to primary colours, and building complex emotions as a combination of the basic ones; this view is in contradiction with the evolutionary notion of basic emotions, since it assumes the possible coexistence of several basic emotions at the same time.

In the following subsections we expose some theories that subscribe to each of these trends.

1.2.1.1 Evolutionary theories

In a context where emotions were mainly investigated by philosophy, and considered irrational, Darwin (1872) proposed the first modern theory of emotions. Several researchers followed him and defined other evolutionary theories, sharing the hypothesis that we humans inherited a small set of adaptive basic emotions along evolution. These theories thus subscribe to the first trend for the definition of basic emotions. Nevertheless they differ on which emotions they consider basic. These theories assume that emotions are innate reflexes, with no cognitive component. We propose a description of some of these evolutionary theories.

Darwin. Darwin (1872) is interested in the selection during evolution of some *basic* emotions for their adaptive function. He studies photographes and uses questionnaires to find a continuity in behaviour and emotional expressions from animals to humans. He then builds a taxonomy associating a specific behaviour and expression to each of his six basic emotions: happiness, sadness, fear, disgust, anger, surprise. He shows that emotional behaviours have an adaptive function, *viz*. they help the individual to face the dangers in his environment and to survive. Besides, he believes that emotions also have a communicative function, since their expression serves to communicate the individual's intentions (for example an animal bares his teeth to indicate his intention to attack). These facial expressions still last now because they are useful to communicate one's emotions to others.

Ekman. Ekman (1992b; 1992a; 1999a; 1999b) also agrees on the existence of a limited set of basic emotions, that are present in other primates and each have specific feelings, universal signals, and corresponding physiological changes. He also agrees with the evolutionary thesis assuming that emotions are inherited adaptive functions.

> "Yet I believe the primary function of emotion is to mobilize the organism to deal quickly with important interpersonal encounters, prepared

Table 1.2: Summary of Darwin's theory

Emotions described	Happiness, sadness, fear, disgust, anger, surprise
Why basic	Inherited through evolution
Triggering	Innate reflex
Discrete emotions	Yes
Causality	Emotion → physiological changes
Cognitive component	No

> to do so by what types of activity have been adaptive in the past. The past refers in part to what has been adaptive in the past history of our species, and the past refers also to what has been adaptive in our own individual life history." (Ekman, 1999a, p. 2)

According to him, emotions are triggered by the automatic appraisal of universal antecedent events and are characterized by their quick onset, their brief duration, and their unbidden occurrence (see all his characteristics of basic emotions in (Ekman, 1999a, table 3.1 p.9)). Then these emotions cause the modifications of facial expression. This assumption is contrary to the facial retroaction thesis adopted by Tomkins and Izard (*cf.* Section 1.2.1.2 below).

Ekman distinguishes six of these basic emotions whose expression is universal, even if it can be inhibited or highlighted by social and cultural rules. These expressions are supposed to have had an adaptive function:

- **joy** is expressed by a true smile, involving a spontaneous contraction of the eye orbicular muscle, the rest of the body being relaxed;
- **disgust** implies a universal grimace which reminds the action of spiting out a poison or pinching one's nose to avoid a bad smell (according to Darwin);
- **surprise** is characterized by eyebrow rising, giving more light to eyes to facilitate the perception of potential threats, it is a state of acute arousal useful to face unknown situations;
- **sadness** is expressed by the relaxation of jaw and the contraction of eyebrows, accompanied by a desire of withdrawal and loneliness, favorable to soothe grief;
- **anger** is accompanied by an attack preparation expression, a quick mobilisation of energetic resources, and a blood influx in hands to make them ready to move;

- **fear** makes the face relax and turn pale, eyes suddenly open to increase the visual ability, and blood floods in legs to favor flight.

Later, Ekman studied the influence of the contraction of each facial muscle on the facial expression, and designed the Facial Action Coding System, a system to describe the facial behaviour in terms of Action Units (Ekman and Friesen, 1978; Ekman, Friesen, and Hager, 2002). Action Units are kinds of atomic facial modifications, corresponding to the contraction of part of a muscle, or of one or several muscles at the same time. He then proposed a correspondence between the felt emotion and the Action Units involved in its expression.

Table 1.3: Summary of Ekman's theory

Emotions described	Joy, sadness, disgust, surprise, anger, fear
Why basic	Universal facial expression
Triggering	Automatic appraisal of universal antecedent events
Discrete emotions	Yes
Causality	Emotion → physiological changes
Cognitive component	No

Oatley and Johnson-Laird. Oatley and colleagues (Oatley and Johnson-Laird, 1987; Oatley, 1992; Johnson-Laird and Oatley, 1992; Oatley and Jenkins, 1996; Oatley and Johnson-Laird, 1996) consider that emotions play a crucial adaptive role in the organization of cognitive processes. They believe that this role has been shaped by evolution.

> What [emotions] do is prompt us, create an urge and a readiness, to act in a way that on average, **during the course of evolution** and assisted by our own development, has been better either than simply acting randomly or than becoming lost in thought trying to calculate the best possible action.
> (Oatley and Jenkins, 1996, p. 261) (bold is mine)

Their main postulate is that emotions play a communicative role.

> "Each goal and plan has a monitoring mechanism that evaluates events relevant to it. When a substantial change of probability occurs of achieving an important goal or subgoal, the monitoring mechanism broadcasts to the whole cognitive system a signal that can set it into readiness to respond to this change. Humans experience these signals

and the states of readiness they induce as emotions."

(Oatley, 1992, p. 50)

The authors assume that cognitive processes work parallely and asynchronously and communicate with each other in two ways: **the propositional or symbolic communication** allows to share information about the environment; **the non propositional or emotional communication** does not transmit information but interrupts all the processes and shifts them into the *emotional mode*, what increases their attention.

> "Emotion signals provide a specific communication system which can invoke the actions of some processors [modules] and switch others off. It sets the whole system into an organized emotion mode without propositional data having to be evaluated by a high-level conscious operating system... The emotion signal simply propagates globally through the system to set into one of a small number of emotion modes." (Oatley and Johnson-Laird, 1987, p. 33)

Since these basic emotions have no propositional content, Oatley *et al*. assume that no high-level conscious processing is involved.

Emotions are triggered by particular "plan junctures" *viz*. particular situations of the current plans. As the other evolutionary authors, Oatley *et al*. propose that emotions then have adaptive effects, here expressed in terms of modifications on the agent's plans. Thus emotions are control signals monitoring current plans to signal problems in their execution. Once triggered each emotion also induces specific physiological changes. Table 1.4 describes the possible emotions modes, along with their corresponding eliciting situations and their adaptive effect on the agent's plans.

Table 1.4: Oatley et al.'s table of basic emotions (Oatley, 1992, p. 55)

Emotion	**Elicitor: situation of current plan**	**Effect on current plan**
Happiness	Subgoals achieved	Continue plan, modify it if needed
Sadness	Major plan failed	Do nothing or find new plan
Anxiety	Self-preservation goal violated	Stop current plan, monitor environment, keep posted or flee
Anger	Active plan frustrated	Try again, attack
Disgust	Gustatory goal violated	Reject substance, withdraw

One particularity of this theory is that it is computationally tractable. Indeed the authors noticed that computational models of human mind and reasoning neglect emotions and thus decided to fill this gap by proposing this *Communicative Theory of Emotions*.

Table 1.5: Summary of Oatley et al.'s theory

Discrete emotions	Yes
Emotions described	Anger, disgust, anxiety, happiness, sadness
Why basic	No required propositional content
Triggering	Modification of plan juncture
Causality	Emotion → physiological changes
Cognitive component	No high-level evaluation
Oriented towards AI	Yes

Conclusion. Evolutionary theories highlight the communicative and adaptive functions of emotions. They agree on the fact that emotions are inherited reflex and thus need no cognition. They also agree on the emotional sequence: emotions trigger (adaptive) physiological changes. This sequence is contradicted by other discrete theories exposed in the next section.

Table 1.6: Summary table: evolutionary theories

Basic emotions	Yes
Why basic	Adaptive emotions inherited through evolution
Basic emotions	No agreement on number or names
Causality	Emotions → (adaptive) physiological changes
Cognitive component	No

1.2.1.2 Well-differentiated discrete emotions *vs* continuous emotions

Discrete theories also consider that emotions are well differentiated from one another, but they do not build this assumption on an evolutionary view. These theories are opposed to continuous ones assume that emotions all are similar phenomena, only different by the values of two or three dimensions. The following paragraphs describe two discrete theories that do not take an evolutionary view, and a continuous one.

Tomkins. Tomkins (1962, 1963, 1980) is interested in primary emotions that he calls affects. They consist in organized physiological, bodily and facial responses to stimuli, directly triggered by neural activation without any cognitive processing. Actually, he defines a notion of density of neural firing (the number of neurons used per time unit) whose modification (either an increase, a decrease, or a maintain at a high level) triggers an emotion. These changes in the density of neural firing can be induced by an innate releaser, a drive, another emotion... without any cognitive appraisal. For example when activation increases, the individual will feel either fear or interest, depending on the suddenness of the change. Neural activation amplifies the physiological responses, then their perception by the individual makes him feel the affect. According to his "facial retroaction" thesis, the facial expression strongly influences the induction and determination of emotion. Tomkins distinguishes nine primary affects having different corresponding facial expressions: interest, joy, surprise, anxiety, fear, and anger, that are innate, and contempt, disgust, and shame, that are acquired. For example surprise corresponds to lifted eyebrows and blinking eyes.

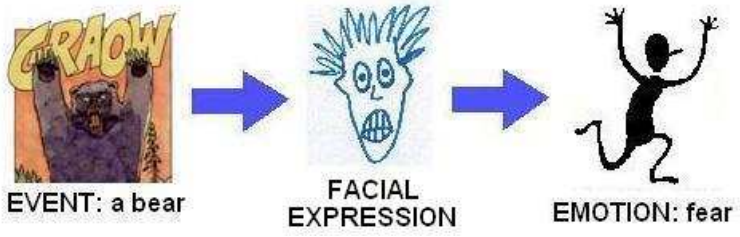

Figure 1.1: Tomkins' facial retroaction hypothesis

Tomkins thus assumes that physiological changes (here the facial expression) precede the feeling of an emotion. This sequence has been subject to debate within the branch of physiological theories (*cf.* Section 1.2.2).

Izard. Izard (1977; 1992) gets his inspiration from Tomkins and even worked with him (Tomkins and Izard, 1965). Like him he insists on the motivational and adaptive significance of affects. His model, named "Differential Emotions Theory" because of its focus on distinct discrete emotions is a model of the interplay between emotion and other personality subsystems[1]. He distinguishes ten fundamen-

[1]According to Izard, personality is constituted of six autonomous subsystems, independent but complexly interrelated: emotions, drives (information signals about bodily needs, like hunger, thirst, reproduction, pain avoidance, comfort), homeostatic (automatic and unconscious systems, like the

Table 1.7: Summary of Tomkins's theory

Emotions described	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Why basic	density of neural firing
Triggering	Facial retroaction
Discrete emotions	Yes
Causality	Physiological changes → emotion
Cognitive component	No

tal emotions (Tomkins' nine emotions more one emotion of guilt) differentiated by specific values of their three components:

- a subjective component *viz.* the emotion consciousness or feeling;
- a neurophysiological component *viz.* innately stored neural programs;
- an expressive component *viz.* characteristic and universally understood (facial, vocal, gestural, and physiologic) observable expressions.

Thus each emotion is triggered by a specific neural activation (induced by internal or external events), then induces in the individual a specific experience and expression, and has a specific influence on his whole behaviour, including perceptions, cognition, action, and personality development. The main assumption of this theory is that "emotions constitute the primary motivational system of human beings" (Izard, 1977, p. 1). Moreover Izard explicitly states that:

> "emotion has no cognitive component. I maintain that the emotion process is bounded by the feeling that derives directly from the activity of the neurochemical substrates." (Izard, 1984, p. 24)

Izard also argues that "the emotions play an important role in organizing, motivating and sustaining behavior" and that "each of the fundamental emotions has an inherently adaptive function" (Izard, 1977, p. 83). Finally, according to Izard, basic emotions are complex phenomena that motivate and organize all the behaviours in an adaptive goal. Past experience shape the future behaviour. Thus emotions are neither disorganized nor disorganizing, but adaptive and functional (in developing and regulating relationships and communications).

cardiovascular one), perceptual, cognitive and motor.

Table 1.8: Summary of Izard's theory

Emotions described	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise, guilt
Why basic	Innate neural program
Triggering	Specific neural activation
Discrete emotions	Yes
Causality	Physiological changes → emotion
Cognitive component	No

Russell. The continuous approach assumes that emotions can be represented with a few dimensions in a multidimensional space. Most researchers (*e.g.* Lang (1994); Russell (1980)) agree on two dimensions: valence is the intrinsic pleasantness of the emotion, differentiating positive emotions like joy from negative ones like anger; arousal is the bodily activation concomitant with the emotion, manifest through physiological changes like increasing heart rate, perspiration... Besides, some researchers argue that two dimensions are not sufficient: for example, anger and fear have the same valence and arousal while they are very different emotions. They thus add a third dimension, named control, potency, or dominance (Russell, 1997), allowing to differentiate emotions w.r.t. the associated action tendency, either fight or flight.

Table 1.9: Summary of continuous theories

Emotions described	Continuous function of arousal, valence and dominance
Triggering	Computation of the values of dimensions depending on stimulus
Discrete emotions	No

Conclusion. There is thus a debate on the representation of emotions. Table 1.10 summarizes the views of these two non-evolutionary discrete theories.

1.2.1.3 Emotions as building blocks

Plutchik. Plutchik (1980) designed the "psychoevolutionary theory of emotions", representing emotions on a coloured wheel. He differentiates eight primary emotions (acceptance, fear, surprise, sadness, disgust, anger, anticipation/curiosity, and

Table 1.10: Summary of non-evolutionary discrete theories

Basic emotions	Yes
Why basic	Biologically differentiated
Basic emotions	No agreement on number
Causality	Physiological changes (facial expression) $\rightarrow$ emotions
Cognitive component	No

joy) that can blend like primary colors to form secondary emotions. His argument to differentiate these eight emotions is that they correspond to specific adaptive processes (safety, reproduction, socialization...). He also specifies the triggering situations of these emotions, and the resulting adaptive behaviour or expression.

- **fear** is elicited by a threat or danger; it results in a behaviour of escape or flight aiming at seeking for protection and safety;
- **anger** is elicited by an obstacle or enemy, and results in an attack to destroy this obstacle;
- **joy** is elicited when the individual gains possession of a valued object; as a consequence he tries to retain the new resources;
- **sadness** emerges from the loss of a valued object or from an abandonment; as a consequence the individual cries to try to reintegrate the group or recover the object;
- **acceptation** is triggered by the belonging to a group, by having friends; it expresses incorporation and mutual support;
- **disgust** is triggered by disgusting objects or poison; as a consequence the individual vomits the poison or rejects the object;
- **anticipation** (what other authors call hope, curiosity or expectation) arises from the presence of a new territory; it results in a behaviour of exploration aiming at mapping this unknown territory;
- **surprise** is triggered by unexpected events; it results in an interruption of current actions: the individual stops in order to use all his time and resources for reorientation.

Plutchik represents emotions with three dimensions (*cf.* Figure 1.2): their similarity to each other (corresponding to their proximity on the wheel), their positive or negative polarity (the eight emotions are grouped into four pairs of polar opposites, like joy and sadness), and their intensity (he uses three words per emotion, corresponding to three degrees of intensity).

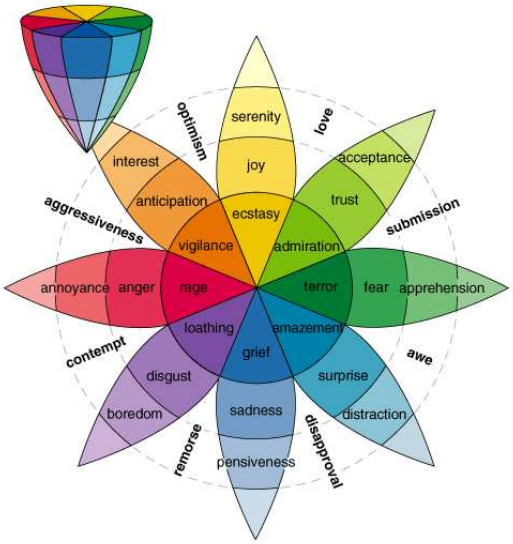

Figure 1.2: Plutchik's circumplex model of emotions

Plutchik assumes that emotions can be felt by all animals, but that their expression was differentiated along evolution and now varies from one species to another, even if some prototypical patterns can be identified. Like other evolutionary theories, Plutchik argues that there is a limited number of basic primary emotions, of which all other emotions are combinations. According to him, they have an adaptive role, in that they help animals to manage their environment when it is threatening their survival.

Table 1.11: Summary of Plutchik's theory

Discrete emotions	Yes
Emotions described	Acceptance, fear, surprise, sadness, disgust, anger, anticipation, and joy
Why basic	Correspondence with adaptive processes

1.2.1.4 Conclusion

Discrete theories all agree on the existence of a limited set of basic emotions, but they reached no consensus about the nature and number of these basic emotions. The authors even differ in their methods and in their criteria for calling an emotion a basic one (*cf.* Table 1.12). Due to this lack of consensus, some authors

Table 1.12: Basic emotions in various theories (table adapted from (Ortony and Turner, 1990))

Author	**Basic emotions**	**Basis for inclusion**
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
Darwin	Happiness, sadness, fear, disgust, anger, surprise	Adaptive processes
Ekman et al.	Joy, sadness, fear, disgust, anger, surprise	Universal facial expressions
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Oatley et al.	Anger, disgust, anxiety, happiness, sadness	No propositional content

are very doubtful about the very existence of such basic emotions. Ortony and Turner (1990) argued that there are no basic emotions at all, neither from a biological nor from a psychological point of view. As a reaction, three researchers answered them (Ekman, 1992a; Panksepp, 1992; Izard, 1992) by arguing once again in favour of the existence of basic emotions.

Most of these discrete theories mainly ground on universal facial expressions to differentiate basic emotions. But several researchers highlighted the fact that the universal expressions identified were posed expressions, played by actors (Scherer and Sangsue, 1995), and not spontaneous. Besides, some studies prove that even actors, referring to standard prototypical expressions, actually show great variability in their expressions (Galati, Scherer, and Ricci-Bitti, 1997, quoted by (Scherer and Sangsue, 1995)). Ekman (1999b) argues against this criticism. He says that the fact that people recognize posed expressions proves that they might have already seen them in their social life. Moreover, he presents results of a study on spontaneous expressions also supporting his thesis of universal expressions.

Thus, representative theories still disagree on the structure of the emotional

space: are emotions discrete or continuous? Other kinds of theories focusing on other aspects of emotions do not always make explicit their view on the representation of emotions. Another important debate about emotions is the question of their triggering: do emotions have a cognitive component? Does brain intervene in the triggering of emotions? The next sections expose two opposing trends of research: physiological theories (Section 1.2.2) highlight the role of physiological changes in the triggering of emotions while cognitive theories (Section 1.2.3) highlight the primacy of cognition in emotion. Contrary to the previous theories that focus on the representation of emotions, physiological and cognitive theories try to explain their triggering.

1.2.2 Physiological theories

They consider that physiological activation is the only origin of emotions. They are not interested in the continuous or discrete representation of emotions.

1.2.2.1 James and Lange

James (1884) postulates that emotions are triggered by the perception of the physiological changes directly induced by a stimulus. Emotions thus are genetical reflexes, which need no intervention of brain.

> "My theory ... is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion. Common sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect ... and that the more rational statement is that we feel sorry because we cry, angry because we strike, afraid because we tremble ... Without the bodily states following on the perception, the latter would be purely cognitive in form, pale, colorless, destitute of emotional warmth. We might then see the bear, and judge it best to run, receive the insult and deem it right to strike, but we should not actually feel afraid or angry."

Actually the individual notices his physiological arousal and deduces that his body is preparing for a particular situation, for example a frightening situation. He then feels the corresponding emotion.

Lange[2] agrees with James about this conception of emotion as a consequence of the perception of physiological changes. But for him, physiological changes are

[2]His text is originally published in Danish, but it was reprinted, *cf.* (Lange and James, 1967)

not reflex, but are controlled by a part of the brain, the vasomotor center.

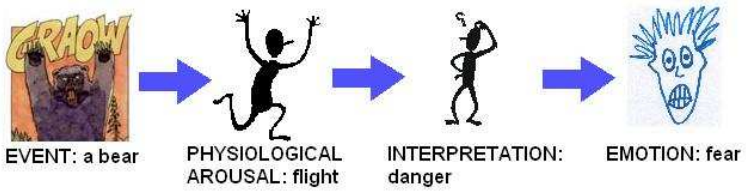

Figure 1.3: James and Lange's emotional sequence

1.2.2.2 Cannon

Cannon (1927) contradicts the main assumption of the two previous theories, *viz.* the fact that physiological responses are differentiated among emotions. According to him viscera react too slowly, so we must feel the emotion before the physiological reaction. Moreover he notices that the physiological changes are quite the same for all emotions, so viscera may not be sensitive enough to allow the differentiation of emotions. He refers to physiological and anatomical experiences showing that artificial induction of physiological modifications is not sufficient to trigger a determinate emotion. He thus postulates that emotions actually result from the activation of the thalamus, that then produces the physiological changes.

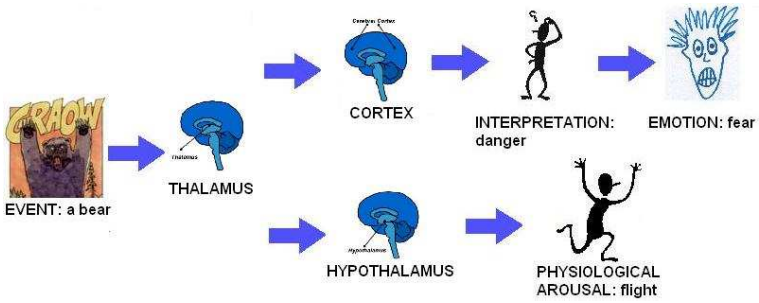

Figure 1.4: Cannon's emotional sequence

But several researchers then showed that Cannon was not right either, since the stimulation of other parts of the brain, like the limbic systems, can also trigger emotions.

1.2.2.3 Conclusion

Even if his hypothesis turned out to be wrong, Cannon was the first one to criticize the physiological differentiation of emotions proposed by James and Lange and also agreed on by Izard or Tomkins who assume that the facial expression determines the emotion. Then Schacter and Singer (1962) also conduced experiments to contradict this facial differentiation of emotions. They injected epinephrin to people in order to artificially induce a physiological arousal. They then noticed that people interpret their arousal depending on the situation: actually some of them were placed in a room with an actor pretending to be happy, others with an actor pretending to be angry, and others were left alone. The results shown that people tend toward feeling the same emotion expressed by the actor, and do not interpret their arousal when they are alone. The authors conclude that physiological arousal alone is not sufficient to trigger an emotion, but that it still has to be interpreted. Actually some emotions can have a distinct biological substrate but a cognitive factor is essential to make subtle differences between close emotions like shame, guilt and embarrassment. This is the basic postulate of the cognitive theories exposed in the next section.

1.2.3 Cognitive theories

The physiological theories assume that stimuli cause a physiological activation (respectively of the viscera, the vasomotor center, or the thalamus) that creates emotions, with no cognitive intervention. The evolutionary theories assume that emotions were inherited during evolution and are automatically triggered by innately stored programs with no cognitive intervention either. On the contrary, cognitive theories assume that cognition is essential in the triggering of emotion.

In this section we present three different kinds of cognitive theories: the first cognitive theories assuming that physiological activation must be interpreted to identify the corresponding emotion (Schacter and Singer, Valins); the cognitive appraisal theories introducing the concept of appraisal, a process of evaluation of the stimuli w.r.t. various criteria (Arnold, Frijda, Lazarus, Scherer, Ortony, Clore and Collins); and the schematic theories assuming that emotions are activated when their cognitive elements are activated (Leventhal).

1.2.3.1 First cognitive theories

Schacter and Singer. Schacter and Singer (1962) agree on the importance of activation but insist on the need to cognitively interpret this activation to identify the emotion that is felt. They thus consider emotions as the result of both physiological

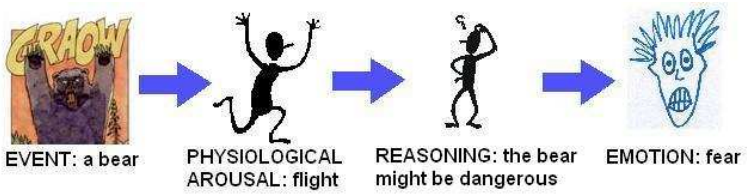

Figure 1.5: Schacter and Singer's emotional sequence

activation and cognition. Their two-factor theory gets its name from their hypothesis that emotions have two components: the physiological arousal of the individual's body, and the individual's cognitive explanation for these changes. This is different from James and Lange's theory assuming that the individual can directly associate a particular kind of physiological arousal with a particular emotion.

They refer to sociological theories to assume that an individual needs to understand his physiological state at every moment. Thus when he experiences an undifferentiated activation, he tries to find a cognition allowing to specify the corresponding emotion. If he has no relevant cognition he will identify his emotion by comparison with other individuals in the same situation. The authors support their theory with an experience where they artificially induce a physiological activation (by adrenaline injection) and then manipulate the resulting emotion by giving the individuals various explanations or putting them in touch with an actor simulating a given emotion (*cf.* Section 1.2.2).

Valins. Valins (1966) is interested in the role of environment in the interpretation of emotions. He completes Schacter and Singer's works by showing that physiological activation itself is not necessary, since the belief of activation is sufficient to create an emotion. Emotion is thus only constituted of two cognitions (the belief of an activation and the causal attribution of this activation to a stimulus).

Finally, the emotional sequence is the same in these first cognitive theories as in James and Lange's theory: the arousal comes first, followed by the emotion. However the difference is that a cognitive appraisal mediates between the arousal and the emotion. Actually, the individual appraises his physiological arousal depending on the context, and does not appraise the stimulus. This is the difference with cognitive appraisal theories.

1.2.3.2 Appraisal theories

These particular kinds of cognitive theories insist on the cognitive determination of emotion and on its adaptive function. They assume that emotions are triggered by the evaluation of a stimulus w.r.t. several criteria: this process is called *appraisal* (*cf*. Figure 1.6). The term "appraisal" was first introduced by Arnold (1960), along with the notion of "action tendency". The following paragraphs present Arnold's theory and some of the main subsequent cognitive appraisal theories.

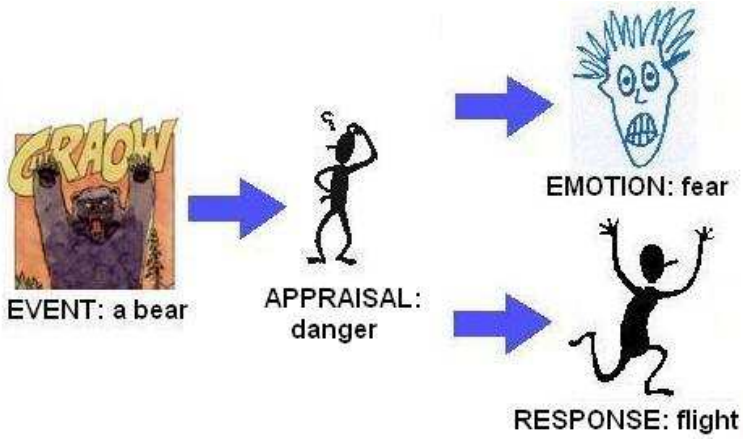

Figure 1.6: The cognitive appraisal theories

Arnold. Arnold (1960) is the one who first introduced the concept of appraisal. According to her, appraisal is the process determining the significance of a situation for the individual, *viz*. is it good or bad for him. This processes triggers an emotion that induces an action tendency of attraction or repulsion, and the corresponding physiological changes. The aim of this readiness is to adapt to the environment.
Later, many researchers agreed on the concepts of appraisal and action tendency introduced by Arnold. We give the broad lines of some of these cognitive appraisal theories in the following paragraphs.

Lazarus. Lazarus (1966; 1991) presents a relational, motivational, cognitive theory of emotion. According to him emotions result from the cognitive appraisal of the interaction between an individual and his environment in relation to the individual's motivations. He distinguishes primary appraisal, evaluating the relevance and congruence of the stimulus to the individual's well-being, and secondary appraisal,

Table 1.13: Arnold's appraisal theory: summary

Automatic appraisal	Yes, automatic, unconscious
Criteria	Good (induces attraction) or bad (induces repulsion)
Discrete emotions	Yes: Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Why basic	Relation to action tendencies
Determinant of emotion	Specific action tendency
Consequences of emotion	Action readiness, physiological changes

evaluating the resources available to cope with this stimulus. Like Arnold, he considers that emotions induce action tendencies and physiological modifications deriving from these action tendencies, with the aim of adapting to the environment.

Table 1.14: Lazarus' appraisal theory: summary

Automatic appraisal	Dual: some appraisals are automatic, others are cognitively controlled
Criteria	Goal relevance, goal congruence, ego-involvement, credit, future expectations, coping potential
Discrete emotions	No
Determinant of emotion	A particular appraisal pattern
Consequences of emotions	Action readiness, physiological changes

Scherer. Scherer (1984; 1987) considers emotions as a multicomponent process, including a cognitive component. His appraisal process consists in a sequence of stimulus processing steps, the "Stimulus Evaluation Checks". They evaluate in turn: the novelty and unexpectedness of the stimulus, its intrinsic pleasantness, its congruence with goals, the coping possibilities, and its compatibility with norms.

This appraisal involves two automatic processes under the control of a conscious process (Leventhal and Scherer, 1987). At the **sensorimotor level**, a first process is responsible for the automatic unconscious perceptual processing of the stimulus w.r.t. innate sensors, and induces reflex responses. At the **schematic level**, a second process is responsible for the automatic unconscious matching of the current stimulus with learned stimulus patterns, and induces coordinated responses.

At the **conceptual level**, a conscious process intervenes when the two automatic ones generate a response that is intense enough to be conscious, and uses propositional knowledge to refine this emotional response. This process becomes more and more automatic by repetition. Scherer and Sangsue (1995) then classify the *Stimulus Evaluation Checks* in these three levels of processing.

Scherer (2001) also proposed a correspondence between emotions and different types of physiological changes. For example facial expression corresponding with an emotion is expressed in terms of Ekman's Action Units.

Table 1.15: Scherer's appraisal theory: summary

Automatic appraisal	Consciously controlled automatic appraisal
Criteria	novelty, unexpectedness, intrinsic pleasantness congruence with goals, coping possibilities, compatibility with norms
Discrete emotions	No
Determinant of emotion	Sequenced evaluation of appraisal criteria
Consequences of emotions	Physiological changes (facial, vocal...)

Frijda. 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). A control signal is then generated to distract or interrupt the current action. An action preparation is then generated (action plan, action tendency, activation mode) that induces physiological changes, and finally an action is selected and executed. According to Frijda it is the associated action tendency that differentiates emotions from each other.

Table 1.16: Frijda's appraisal theory: summary

Criteria	Causes and consequences, relevance and congruence with interests, coping possibilities, urgency
Determinant of emotion	A particular action tendency
Consequences of emotions	Action preparation, physiological changes

Ortony, Clore and Collins. Ortony, Clore, and Collins (1988) considered emotions as being valenced reactions to three kinds of stimuli: events, actions of agents, aspects of objects. They thus design a typology (known as the OCC typology) that has three branches. Each branch corresponds to a type of stimulus appraised with respect to a particular appraisal variable, and related to particular mental attitudes. The **event-based branch** contains emotion types whose eliciting conditions depend on the evaluation of the *desirability* of an event with respect to the agent's goals. For example, the stimulus event "it is raining" is appraised as being undesirable w.r.t. the agent's goal of taking coffee on a terrace. The **agent-based branch** contains emotion types whose eliciting conditions depend on the judgement of the *praiseworthiness* of an action, with respect to the agent's standards. The **object-based branch** contains emotion types whose eliciting conditions depend on the evaluation of the *attraction* of an object with respect to the agent's likings. These branches are then differentiated into several groups of emotion types with similar eliciting conditions, depending on other criteria called *local variables*. Other criteria called *intensity variables* affect the intensity of these emotions.

Table 1.17: Ortony, Clore and Collins' appraisal theory: summary

Criteria	Desirability of an event, praiseworthiness of an action, attraction of an object + local variables
Discrete emotions	No (generic categories of emotion types)
Determinant of emotion	Particular eliciting conditions
Intended for AI applications	Yes

Conclusion. The common assumption in all these theories is that individuals continuously appraise their environment w.r.t. various criteria, mainly its relevance for their well-being. This concept of appraisal was introduced by Arnold, and the subsequent theories bring some differences to it. Lazarus assumes that appraisal is necessary but also sufficient to trigger emotions, and asserts that emotions are thus entirely determined by a particular appraisal pattern. Scherer or Ortony, Clore and Collins also assume a correspondence between appraisal patterns and emotions. On the contrary, Arnold or Frijda believe that what determines emotions is the associated action tendency; Frijda adds new action tendencies to the two ones postulated by Arnold (attraction, aversion). Another common point between these theories is their adaptive point of view: they all agree on the fact that emotions induce action tendencies and corresponding physiological changes aiming at a better adaptation to the environment.

Cognitive appraisal theories and discrete theories were considered to be very different approaches to emotions ((Ortony and Turner, 1990), *cf.* Section 1.2.1.4). However, appraisal theories also focus on the description of a small number of emotions. Actually, the difference is that cognitive appraisal theories do not explicitly assume that these emotions are (biologically, adaptively...) more basic than the other ones.

Table 1.18: Appraisal theories: summary table

Appraisal criteria	No agreement
Discrete emotions	Yes (Arnold), No (Lazarus, Ortony et al), implicit
Sequence of application	Yes (Scherer), No (Lazarus)
Conscious appraisal	No (Arnold), dual (Lazarus), controlled (Scherer)
Determinant of emotions	Appraisal frame (Lazarus, OCC, Scherer) *vs* action tendency (Arnold, Frijda)

1.2.3.3 Schematic theories

Other cognitive theories are the schematic theories. They assume that memory stores some emotional information allowing to reactivate an emotion when one of its elements is activated, even if this emotion has no link with the stimulus (see for example Leventhal 1980 ou 1984).

1.2.3.4 Conclusion

All these cognitive theories highlight the fact that cognition is crucial to determine which emotion the individual feels in a given situation. Some of them (like Valins) even assert that cognition is sufficient to trigger an emotion. This too deliberative view on emotion was much criticized (in particular by Zajonc, see Section 1.1). Then Leventhal and Scherer (1987) showed that this debate was actually about the definition of cognition and proposed to differentiate various cognitive processes at various levels of consciousness. Then many other researchers agreed on this combination of automatic and controlled processing to generate emotions. For example, neuropsychology also distinguishes between innate primary emotions handled by the limbic system, and more complex learned secondary emotions, or social emotions, needing a cortical processing. Some appraisal researchers then refined their model (Lazarus, 1991; Ortony, Clore, and Collins, 1988; Clore and Ortony, 2000) to include this notion of duality: emotional responses are produced by the

interaction of automatic processing (perceptual processing of stimuli and comparison with existing schemas) and conscious effortful processing. Thus finally, there seems to be a consensus that cognition plays a role in emotion, but at different levels depending on the situation.

We will now discuss why, among the various trends of research presented in this review, cognitive appraisal theories are more adapted to reach our goal of formalizing emotions. The next section is dedicated to this trend of research: it exposes its assets and then give more details on two theories of this trend that we will use in the following of the thesis.

1.3 Cognitive appraisal theories: assets and details

1.3.1 Advantages of cognitive appraisal theories over other theories

We have exposed several kinds of theories of emotions, that differ on various features. We will now discuss the assets of appraisal theories in comparison to other theories, as exposed by (Scherer, Schorr, and Johnstone, 2001).

First, cognitive appraisal theories allow a fine-grained differentiation between emotions. The theories exposed before propose various explanations for the differentiation of emotions. Physiological theories explain this differentiation by a specific biological substrate, but experiments show that visceral responses are mainly undifferentiated; continuous theories differentiate emotions by a continuous value of two or three dimensions, but some emotions have quite close values of these dimensions while being subjectively very different; evolutionist theories characterize each emotion by a specific facial expression; some cognitive theories suggest that each emotion is associated with a particular action tendency. But finally all these explanations seem to be insufficient. On the contrary cognitive appraisal theories assume that emotions result from the evaluation of stimuli. Thus each distinct emotion corresponds to a particular appraisal pattern, *viz*. a particular combination of appraisal criteria.

This cognitive mediation between the stimulus and the emotion is also the only way to account for inter-individual differences: in the same situation, several individuals can feel different emotions, that actually result from different evaluations of the situation. It also accounts for the temporal variability of emotions: the same individual can appraise differently the same situation at different moments, and thus feel different emotions about it depending on the current context. On the contrary, a direct correspondence between the stimulus and the emotion cannot account for this variability, since the stimulus itself does not vary. Cognitive appraisal theories also explain cross-situational similarities, *viz*. the fact that a large number of situations (actually an infinity) can trigger the same emotion: all these situations match

the same appraisal pattern. Besides they can be completely new situations, with no concrete common features with other known situations. Theories assuming that emotions are conditioned responses specified by evolution cannot account for this kind of phenomena.

Moreover, the theories focusing on the physiological triggering of emotions, independently from the situation, cannot explain the appropriateness of the emotion to the particular current situation. Now studies show that emotions are not disorganized as it was believed before, but on the contrary have an adaptive function to deal with their triggering situation. Cognitive appraisal theories account for this adaptive role since they propose that emotions are triggered by the evaluation of a particular situation and then induce an appropriate action tendency to deal with this situation.

Finally, we will give significantly more details about the two appraisal theories that are mainly used in computer science: Lazarus' cognitive motivational relational theory, and Ortony, Clore and Collins' typology of types of emotions. The latter is the theory that we will formalize in the sequel of this work.

1.3.2 Lazarus' motivational relational theory

Lazarus (1991) assumes that an individual continuously evaluates the relation with his environment. Emotions are elicited by particular encounters *viz.* particular relations between the individual (his goals, his preferences) and the environment (its constraints, its resources). According to him, the human emotional process is made up of two indivisible processes: appraisal and coping.

1.3.2.1 Appraisal

According to Lazarus, knowledge or beliefs about the world are not enough to trigger an emotion. Another process is necessary to assess the personal signification of this knowledge for the individual: this is the appraisal process. Indeed, knowledge is neutral, objective, while appraisals are subjective and depend on the person's goals; this accounts for the inter-individual variability in emotional responses to the same situation. Actually, the individual continuously modifies his relationship with his environment as he acts on it; environmental feedbacks then lead to reappraisal of the situation and to new emotions. Finally, "emotions are always in flux" (Lazarus, 1991, p. 134).

Appraisals (and reappraisals) are constituted of two complementary types of appraisal. **Primary appraisal** assesses the relevance of the encounter to the individual's well-being. It has three components: goal relevance, goal congruence, and type of ego-involvement.

- *Goal relevance* is the importance of the situation for the individual: *does the situation involve issues about which I care?* If the situation is not relevant to any goal, then it can trigger no emotion.

- *Goal congruence*: if the situation is congruent with one of the individual's goals, *viz.* if it facilitates achievement, then a positive emotion will be triggered; if it is incongruent with some goal, *viz.* if it threatens or impedes its achievement, then a negative emotion will be triggered.

- *Type of ego-involvement*[3] gathers several features of ego-identity and personal commitments, sorted in six categories: self- and social esteem, moral values, ego-ideals, meanings and ideas, well-being of other persons, and life goals. Actually, it represents in which way the agent is personally involved in the current situation.

Secondary appraisal assesses the individual's coping options, *viz.* the actions he may perform and their envisaged effects on the situation. It has three components: blame or credit, coping potential, and future expectations.

- *Blame or credit* are an attribution of responsibility to a person accountable for the situation at hand. If this person had control on what happened, then she can receive blame or credit. *Did someone deliberately provoke this situation?*

- *Coping potential* is an evaluation of how the individual can manage the situation, change it or change his goals, in order to restore a good relationship with his environment. *Can I do something to restore the balance between me and my environment?*

- *Future expectations* represent the expected modifications of the situation if the agent does not intervene: *will the situation turn out right if I do nothing?*

Lazarus' appraisal components match those of other researchers (Frijda, 1986; Roseman, 1984; Scherer, 1984; Smith and Ellsworth, 1985). However, contrarily to others like Scherer, he assumes that their evaluation is not sequential. Primary and secondary appraisal are complementary to determine the significance of the encounter for the individual, but they can be performed in any order (Lazarus, 1991, p. 151).

[3]This seems to be the most abstract and complex concept in Lazarus' theory of appraisal. In our view, this is why this theory is so hard to formalize.

1.3.2.2 Emotional appraisal patterns

Lazarus describes nine negative and six positive emotions, and gives their appraisal pattern *viz*. the corresponding values of their appraisal components. He also gives their definitions, that he calls "core relational themes" (Lazarus, 1991, table 3.4 p.122). Below we quote this definition and give the corresponding appraisal pattern (Lazarus, 1991, Chap. 6 and 7). However we do not repeat for each emotion that the situation is goal relevant.

The negative emotions all arise when the situation is goal incongruent.

- **Anger** is "a demeaning offense against me and mine". The agent's self-identity is damaged or threatened by an agent that receives blame. If a viable attack is possible, with positive future expectancies of the environmental response, then anger is facilitated.

- **Fright** is a "concrete and sudden danger of imminent physical harm", ego-involvement is not relevant; **anxiety** is an "uncertain, existential threat", the agent is involved to protect his ego-identity or personal meanings; these two emotions are close but differ in the kind of threat.

- **Guilt** is "having transgressed a moral imperative", a self disgrace, and the individual is involved to manage this moral transgression; **shame** is "a failure to live up to an ego-ideal", a social disgrace, and the agent is involved to manage this failure. Guilt and shame thus differ on their type of ego-involvement. Favorable coping potential and future expectations reduce guilt or shame.

- **Sadness** is due to an "irrevocable loss"; the individual suffers a loss in any kind of ego-involvement, there is no one to blame, and his coping potential is unfavorable. If his future expectations are positive then sadness is blended with hope, else it leads to despair.

- **Envy** is "wanting what someone else has", another individual possesses something lacking to any type of ego-involvement, his responsibility is not relevant to determine envy; favorable coping potential and future expectations increase envy. **Jealousy** is "resenting a third party for loss or threat to another's affection [or favor]": the individual's ego-involvement is threatened by a lack of affection, and someone else receives blame for taking this affection in a contestable way; favorable coping potential and negative future expectations increase jealousy.

- **Disgust** is "taking in or being too close to an indigestible object or idea

(metaphorically speaking)". Any type of ego-involvement is threatened by this "poisonous idea".

Positive emotions mainly arise when the situation is goal congruent, but some of them are problematic[4].

- **Joy or happiness** is a "reasonable progress towards the realization of our goals". Future expectations must be favorable to set up a favorable background for joy.
- **Pride** is an "enhancement of one's ego-identity by taking credit for a valued object or achievement, either our own or that of someone or group with whom we identify". The type of ego-involvement is an enhancement of esteem[5]. One must receive credit for this enhancement in order to feel pride.
- **Love or affection** is "desiring or participating in affection, usually but not necessarily reciprocated". The type of ego-involvement is a desire for mutual appreciation.
- **Relief** arises when "a distressing goal incongruent condition has changed for the better or gone away". This emotion arises when a goal-incongruent situation changes towards a goal-congruent one.
- **Hope** is "fearing the worst but yearning for better". According to Lazarus it is a problematic emotion since it is felt as positive but arises when the situation is goal-incongruent. Moreover the future expectations must be negative but uncertain.
- **Compassion** is "being moved by another's suffering and wanting to help". It is another problematic emotion, felt as positive but arising in a situation that is incongruent with someone else's goals. There must be no one to blame.

1.3.2.3 Coping

Lazarus considers that a second process is intimately linked to appraisal: coping. Lazarus and Folkman (1984) quote two origins of the concept of coping: the darwinian theory of stress and control in animals, defining coping as acts controlling aversive situations to lower psychological and physiological perturbations; and psychoanalytic ego psychology, defining coping as realistic and flexible thoughts

[4]Lazarus call them problematic since they are subjectively felt as being positive, but arise in a goal-incongruent situation.

[5]Please remind that the type of ego-involvement in anger was a damage of self-esteem. Thus Lazarus opposes pride with anger.

and acts that solve problems and reduce stress. But for Lazarus, what is important in coping is not the result but the efforts, that differentiate coping from automatic adaptive processes like action tendencies or reflexes. He thus gives a new definition of coping:

> "constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person" (Lazarus and Folkman, 1984)

Coping thus has to do with the mastering or minimization of stressful situations. Lazarus then distinguishes two kinds of coping: *problem-focused coping* is oriented toward the management of the problem creating the stress, and is more probable when appraisal indicates a possible solution to this problem; *emotion-focused coping* is oriented toward the regulation of the emotional response to the situation, and is more probable when appraisal indicates no solution to the problem.
Lazarus uses a questionnaire called Ways of Coping that he designed with colleagues (Folkman et al., 1986) to determine the coping strategies used by people in stressful situations. He identifies eight strategies.

- **Confrontive Coping** consists in making aggressive efforts and possibly taking risks in order to modify the stressing situation.
- **Distancing** consists in making cognitive efforts to minimize the importance of the situation and detach from it.
- **Self-Controlling** consists in trying to regulate one's emotions and actions.
- **Seeking Social Support** consists in asking for information, material help or emotional support.
- **Accepting Responsibility** consists in admitting one's responsibility in the situation and trying to repair it.
- **Escape-Avoidance** consists in trying to escape or avoid the problem, through wishful thinking or behavioural efforts.
- **Planful Problem Solving** consists in focusing on the problem and deliberating to try to solve it.
- **Positive Reappraisal** consists in trying to find a positive aspect in the situation by focusing on personal growth or religion.

Notice that all these strategies involve conscious efforts from the individual. Emotions are also associated with action tendencies that are innate dispositions, unconscious reflexes involving no effort. On the contrary coping strategies are more complex and deliberate and can inhibit or redirect the action tendency to match the individual's plans.

1.3.2.4 Conclusion

Lazarus' theory is quite complex, particularly from a computer science point of view. In particular the appraisal patterns are quite subjective: emotions correspond with qualitative abstract values of the appraisal components (for example ego involvement). Moreover some appraisal patterns match no emotion while several different appraisal patterns can match the same emotion. Some emotions are also very close and their appraisal patterns only differ on minor points.

Because this theory is so abstract and complex, it is difficult to formalize it in order to implement it. Actually, as far as we know, this theory was only implemented by Gratch and Marsella (2004a) in their EMA agent.

The next section describes another cognitive appraisal theory that is far less complex than Lazarus' one.

1.3.3 Ortony, Clore and Collins' typology of emotions

Ortony, Clore and Collins wanted to develop a cognitive theory of the origin of emotions, and not of their behavioural or physiological consequences. Their aim was to identify classes of emotions gathering several emotion words actually accounting for the same emotion with different intensities or consequences.

Their theory is a cognitive appraisal theory, since the origin of emotions is believed to be the appraisal of an antecedent situation w.r.t. some appraisal variables. Emotions are valenced reactions to three types of stimuli: events, agents, and objects. The theory is structured in three branches each one differentiated in several groups of emotions, depending on the involved appraisal variables. This structure is summarized in Figure 1.7. The first branch gathers emotions resulting from the appraisal of the desirability of an event w.r.t. the agent's goals. The second branch gathers emotions resulting from the appraisal of the praiseworthiness of another agent's action w.r.t. the agent's standards. The third branch gathers emotions resulting from the appraisal of the appealingness of an object w.r.t. the agent's likings. There also exist composed emotions resulting from the conjoint appraisal of several criteria.

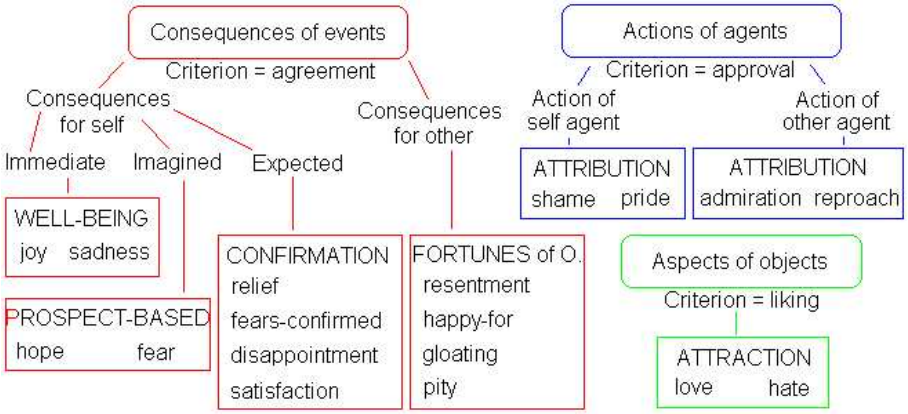

Figure 1.7: The three branches of the OCC typology, corresponding with the three types of stimuli that can trigger an emotion

1.3.3.1 Appraisal variables

Desirability, praiseworthiness, and appealingness are central appraisal variables, that intervene in the determination of the type of emotional reaction to the situation. An event is desirable (resp. undesirable) if it facilitates (resp. interferes with) the agent's main goal, or a subgoal supporting it; goals represent what the agent wants to be. An action is praiseworthy (resp. blameworthy) if it upholds (resp. violates) some of the agent's standards; standards represent what the agent thinks to be normal, *viz.* legal, moral... An object is appealing (resp. unappealing) if it matches (resp. does not match) the agent's likings. These central variables also affect the intensity of the emotions in the corresponding branch: they are local intensity variables. Other local intensity variables are specific to each emotional group, so they will be detailed in the next sections.

Some global intensity variables also affect the intensity of all emotions: *sense of reality* measures how much the emotion-inducing situation is real; *proximity* measures how much the individual feels psychologically close to the emotion-inducing situation; *unexpectedness* measures how much the individual is surprised by the situation; *arousal* measures how much the individual was excited before the stimulus.

1.3.3.2 Reactions to events

The emotion types in this branch all have an eliciting condition and an intensity depending on the desirability of an event w.r.t. the individual's goals. They are divided into four groups of emotions:

- Well-being emotions (joy, distress) correspond to the appraisal of the desirability for self of an event that has occurred.
- Prospect-based emotions (hope, fear) correspond to the appraisal of the desirability for self of an event that could occur.
- Confirmation emotions (satisfaction, fears-confirmed, disappointment, relief) correspond to the appraisal of the desirability for self of an event that was expected before it occurred (these emotions thus arise after a corresponding prospect-based emotion).
- Fortunes-of-others emotions (happy-for, sorry-for, gloating, resentment) correspond to the appraisal of the presumed desirability for another agent of an event.

The intensity of prospect-based and confirmation emotions is also affected by: *likelihood* which is the degree of belief that the prospected event will happen; *effort* which is the degree of utilization of resources to make the prospected event happen or to prevent it from happening; and *realization* which is the degree to which the prospected event actually happens.

The intensity of fortune-of-others emotions is also affected by: *desirability for other* measuring how much the appraised event is presumed to be desirable for the other agent; *liking* indicating if the individual appraising the situation has a positive or negative attitude toward the other individual; *deservingness* measuring how much the individual appraising the situation believes that the other individual deserved what happened to him.

1.3.3.3 Reactions to agents' actions

The emotion types in this branch all have an eliciting condition and an intensity depending on the praiseworthiness of an action w.r.t. the agent's standards. They are structured in one group of emotions. Attribution emotions (pride, shame, admiration, reproach) correspond to the appraisal of the praiseworthiness of an agent's action (who is either the agent himself or another one).

The intensity of attribution emotions is also affected by other parameters: *strength of cognitive unit* measures how much the individual identifies with the individual

or institution that is the author of the action inducing the emotion; *expectation-deviation* measures how much the individual's action deviates from what is usually expected from him, depending on the norms, his social role, his personality...

1.3.3.4 Reactions to objects

The emotion types in this branch all have an eliciting condition and an intensity depending on the appealingness of the aspects of an object w.r.t. the agent's likings. They are structured in one group of emotions. Attraction emotions (love, hate) correspond to the appraisal of the appealingness of the aspects of an object.

The intensity of attraction emotions is also affected by *familiarity*, measuring how much the object is familiar to the individual.

1.3.3.5 Composed emotions

Moreover, there is a group of composed emotions corresponding to the simultaneous appraisal of the situation w.r.t. several central variables: desirability and praiseworthiness. Well-being and attribution composed emotions (gratification, remorse, anger, gratitude) correspond to the appraisal of the consequences of an action and of the praiseworthiness of its author.

1.3.3.6 Conclusion

This theory is finely structured and the values of appraisal variables are simpler than in Lazarus' theory, making it easier to formalize. This may result from the authors' intention of proposing a computationally tractable theory for AI applications. This goal was quite reached since this typology has already been implemented in a huge number of agents (we give some examples in the next chapter), what shows that it is also well adapted to their designers' goals. For all these reasons it is this theory that we choose to formalize in this work.

1.4 Conclusion

The study of emotions is an active field of research in philosophy, biology, psychology and sociology. This shows that emotions are a complex phenomenon involving various aspects: physiological modifications, innate reflexes, subjective feelings, cognitive evaluation, impact on social interactions... In spite of many debates between these various views, there seems to be a consensus on the role of cognition in the triggering of emotions and on their adaptive function. We also show that

cognitive appraisal theories, assuming that emotions are triggered by the evaluation of stimuli w.r.t. various criteria, have many assets and are more adapted to be formalized.

This chapter was intended to answer the question of **what is an emotion?** Our historical review shows that there are several opposite views. Actually, various theories often focus on various aspects of emotions and answer different questions about them. So finally the existing theories are both contradictory on their answers to questions and complementary because they do not answer the same questions. The question in which we are more interested here is the cognitive triggering of emotions. Cognitive appraisal theories all answer this question in a rather close way, but we need to choose one of them to set up the basis of our model. Indeed to endow an agent with believable emotions it is important to build on psychological definitions that have already been experimented for a long time. Actually, few of the theories reviewed in this chapter are both computationally tractable and explicative enough of the triggering of emotions. Finally we choose to formalize the OCC typology, for various reasons. First, the authors claim that their theory is intended for Artificial Intelligence applications. As a consequence of this choice, their theory is very structured and uses a limited number of quite comprehensible concepts, what makes it easier to understand even for a computer scientist. On the contrary Lazarus uses some highly abstract criteria like ego-involvement, which appeared hard to formalize. Second it is already well structured in several categories of emotions with similar eliciting conditions, what simplifies our work. Moreover the OCC typology seems to cover quite exhaustively the range of situations that an artificial agent can face. Of course at this point, and in spite of all our efforts of comparison, we cannot say that one theory is better than the other ones. Nevertheless the fact that most of existing emotional agents ground on this OCC typology gives it a kind of legitimacy: it seems to be adapted to the aim of most of emotional agents designers, what tends to prove its validity as a basis for our model. To further convince the reader of our choice, we will discuss this choice again in more details later: when formalizing emotions (in Chapter 4) we will compare Ortony *et al.*'s sayings with those of Lazarus on each formalized emotion. Moreover our formalization will allow us to assess the OCC typology of emotions (*cf.* Chapter 7).

Once this choice is made (even if it is still debatable for now), we can address the second question: **why endow virtual agents with emotions?** Actually we saw that several psychologists are already convinced of the crucial role emotions will play in intelligent systems, and propose computationally tractable models of emotions: Oatley and Johnson-Laird (*cf.* Section 1.2.1.1), and Ortony, Clore and Collins (*cf.* Sections 1.2.3.2 and 1.3.3). The role of emotions in human intelligence is also supported by recent findings made possible by advances in neuroscience and

medical imagery (Damasio, 1994). Yet meanwhile computer scientists still exclude emotions from the design of rational intelligent agents. Agent designers have long been incredulous that emotions, even if crucial for humans, are also needed for their agents: virtual agents may not need to function as we do. But computer science recently began, with success, to endow agents with emotions, mainly by implementing (part of) the OCC typology.

It is important to understand that our aim is not to design another emotional agent implementing the OCC typology. This thesis rather proposes to design a formal model of emotions, intended to serve as a generic reusable basis for future implementations in various agents for various applications. The next chapter, providing a review of existing work about emotions in computer science, is thus intended to convince the reader of the crucial role that emotions play in this field, and then of the usefulness of our model.

Chapter 2

Emotions in computer science

The question is not whether intelligent machines
can have emotions, but whether machines can
be intelligent without any emotions.
(Minsky, 1988)

2.1 Introduction

In this chapter we get interested in the use of the psychological theories of emotions in a field that may seem completely opposite: computer science. After having neglected emotions for decades, computer science starts getting interested in their potential for artificial agents in the 90s. It is important to notice that what is called emotions for an agent does not match exactly human emotions. As Picard (1997) says, they are rather labels characterizing a particular mental state.

> The computer's emotions are labels for states that may not exactly match the analogous human feelings, but that initiate behaviour we would expect someone in that state to display.
> (Picard, 1997, p. 298)

In the Section 2.2 we discuss some early work from various fields of research aiming at integrating emotions into virtual agents or at proving their usefulness for such agents. Then we describe some emotional architectures (Section 2.3) and various application fields where they are used increasingly often: pedagogical agents (Section 2.4) and conversational agents (Section 2.5). We then notice that all these applications directly implement emotions into their agents. These emotional models may thus be *ad hoc*, specific to their application, not generic and not reusable.

Therefore a new line of research emerges, aiming at designing generic formal models of emotions. In Section 2.6 we discuss why BDI logics can be used to design such models, and describe some attempts in this sense.

2.2 Why use emotions in computer science?

The previous chapter showed that psychology have now acknowledged the role of emotions in the functioning of human mind. Several theories recognize their adaptive function (*cf.* Chapter 1). Moreover some psychologists are even interested in providing computationally tractable models in order to allow AI researchers to integrate emotions into their intelligent systems. Oatley and Johnson-Laird, as well as Ortony, Clore and Collins, were early interested in the role of emotions in computer science systems, and their theories are intended for AI applications.

But meanwhile computer scientists, even if they start looking at emotions, are only interested in making their agents more believable. They do not care yet about the functions of emotions and their influence on behaviour and social interaction. This notion of "believability" was first introduced by Bates (Section 2.2.1). Then several other fields of research, like neurology, also support the role of emotions in intelligence and design models of their impact on reasoning (Section 2.2.2). Some computer scientists then start getting aware of the importance of emotions not only for their agent's believability but also for their impact on interactions with the user: this is the beginning of Affective Computing (Section 2.2.3). Emotional agents become fashionable and various works appear. We then summarize Gratch and Marsella's attempt to clarify emotional psychology for agents designers (Section 2.2.4).

2.2.1 The role of emotions for believable agents

Bates' 1994 central assumption is that AI researchers wanting to create lifelike characters should find insight in artistic work about "believable characters", in particular in the field of animation. By analogy with it, he introduces the term of "believable agents" that he defines in the following way:

> "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. 1)

He notices that AI researchers try to create the illusion of life by endowing agents with human abilities that they consider essential in intelligence (like reasoning, problem solving, learning...), but then rather create artificial scientists than artificial humans. On the contrary, cartoon animators have to abstract the very essence

of human life in order to give life to simplistic unrealistic drawings. In particular, animators highlight that one crucial quality of believable characters is to have "appropriately timed and clearly expressed emotion". Thus when representing the emotions of a character they obey three important rules. First the character's emotional state must always be clearly defined in order to be clearly identifiable. Second, this emotional state must be reflected by the character's actions and reasoning. Third, animators must use various techniques to help the spectator decode the expressed emotion, like stylizing or exaggerating it. Bates builds on these rules to create believable agents. In the Oz project, Bates (1992) designs a virtual world inhabited by Woggles, animated emotional creatures.

First, the Woggles' emotional model is built on the OCC typology, which guarantees that these creatures always have a well-defined emotional state. Second, each emotion is mapped with a "behavioural feature" that influences the subsequent behaviour. Bates' Woggles then meet the first two requirements of believable characters. Nevertheless Bates does not use any technique to make their emotional state easily decodable. These techniques, like exaggerating the agent's emotions, are unrealistic but crucial for spectators to understand the scene. Finally, artists and AI researchers have the same goal, so Bates conclude that AI should find insight in cartoon animators' work.

Other researchers show this central role of emotions and personality in believable agents (Rousseau and Hayes-Roth, 1998). Then several research findings help showing that emotions do not only improve believability but also impact cognition.

2.2.2 Some proofs of the emotional impact on cognition

The neuroscientist Damasio (1994) conduced well-known experiments showing that a patient with brain damages preventing him from feeling emotions was also subsequently unable to make decisions or to interact socially. He then formulates the "Somatic Marker Hypothesis" saying that emotions guide and improve human decision-making.

Psychologists also propose models of the impact of emotions on cognitions. Forgas (1995) proposes the *Affect Infusion Model*, a framework allowing to explain how affectively loaded information can modify reasoning. This influence intervenes in two aspects: not only the informational content of cognitions is modified, but also the reasoning processes applied to these cognitions. Indeed the author supposes that individuals dispose of a set of possible reasoning strategies among which they choose depending on the context in order to minimize their efforts. In his model, Forgas proposes four strategies of information processing: *direct processing* retrieves a reaction already triggered in the past; *motivated processing* is a selective research, directed by a motivation towards a precise target; *heuristic*

processing uses a limited set of information and associations to produce a low-cost answer when the two first strategies do not work; *substantial processing* is only used for complex new tasks.

Actually, numerous research findings attest the impact of emotions on all abilities characteristic of human intelligence (planning, memory, learning...). This is not the object of this thesis to exhaustively expose all these works here (see (Gratch and Marsella, 2005) for a review).

2.2.3 The birth of Affective Computing

As a result of the growing number of proofs of the impact of emotions on interaction, Picard (1997) assumes that some computer systems may need to be able to recognize the user's emotions. Nevertheless she concedes that it is a difficult task for a machine since even we who are better informed than anyone about our own emotional state are not always able to recognize or label it.

Picard, Vyzas, and Healey (2001) then developed a technique to measure a user's emotion. Their system is set up of sensors measuring several physiological indicators of the autonomic nervous system changes. They used it to measure the emotional state of a person over a long period of time. Their results contradict Schacter and Singer's assumption (*cf.* Section 1.2.3.1 page 41) that physiological signals are mainly undifferentiated. Indeed Picard and colleagues were able to accurately differentiate eight emotions from their physiological signals. But they suffer from one important limitation: they force the choice between a limited set of eight emotions; only inside this set can the machine accurately determine which emotions is expressed.

Picard (1999) is more particularly interested in Affective Computing for the design of intelligent human-computer interfaces. She believes that Affective Computing improves interfaces by allowing affective communication with the user, and providing ways to exploit the received affective information. Mainly, interfaces can reduce the user's frustration by helping him express his emotions and then recognizing his emotional expressions. Picard highlights that it is important to let the user choose if he wants to express his emotion or not: the user wants to keep control over his expression.

Finally, emotions have been too long neglected in computer systems and researchers must now take them into account. Nevertheless they must use emotion in an intelligent and balanced way: not all computers need emotions. Moreover we can notice that Picard's method to measure the user's emotion is quite intrusive: it reveals that the perception of the user's emotions really improve the system, but this perception should then be done through other means. We will envisage later a solution using the model designed in this thesis (*cf.* Section 6.2).

2.2.4 Useful emotional abilities for virtual agents

Gratch and Marsella (2005) agree on the crucial role emotions play in human intelligence and on the need to integrate them into virtual agents. However they notice that computer scientists are mainly interested in making their agents more believable by endowing them with emotional expressions. They thus want to highlight that emotions are not only a means to make agents more believable and convincing (for example in human-computer interfaces). Indeed they expose several effects of emotions on cognition and interaction, that must be known by agent designers since they can impact their application in a positive but also in a negative way.

The authors identify two main views on the functions of emotions: the intra-agent view interested in their influence on cognitive processes; and the inter-agents view studying their impact on social interaction. They then discuss some generally accepted cognitive functions of emotions. Emotions favor perception and categorization of relevant stimuli; they suggest adapted reactions to the environment; they trigger the use of adaptive coping strategies, unfairly considered to be irrational; they facilitate learning and recall[1].

They also expose some of the generally accepted social functions of emotion. First, emotions (sincere or conventional) can be interpreted to deduce information about the agent's mental state; they communicate these mental states more efficiently than speech. Second emotions can be used to manipulate the hearer's emotions, motivations and behaviour. Third, emotional display makes people more believable and calls for trust and empathy.

Finally agent designers must formalize these functions in order to faithfully simulate social interaction. The authors distinguish between two implementation approaches. **Communication-driven approaches** select an emotional expression because of its communicative or manipulating effect, while **simulation-based approaches** try to simulate the emotional process and to give *true* emotions to the agent. There exist some mixed approaches: for example for Prendinger and Ishizuka (2001), appraisal triggers emotions among which a communicative filter then selects the one that will be expressed to match the dialogue goal.

Finally Gratch and Marsella discuss the problem of the expression of emotion. Should an expressive agent rather express a realistic emotion, through a multimodal behaviour faithful to how humans express their emotions? Or should he on the contrary display stylized and thus unrealistic emotions in order to make it easy for the user to recognize them? We said before that Bates consider the second solution to be better, since humans are not very good at understanding emotions.

[1] Bower (1991) shows that memorized information is affectively loaded depending on the current mental state; then it is easier to retrieve information that is loaded with the same emotion than the current one. This makes it easier to retrieve information that is relevant in the current context.

The central role of emotion in the design of intelligent virtual agents is thus well recognized now: emotion makes these agents more believable, and impacts their cognitive processes and their interaction with the user. Most of computer scientists are now aware of the usefulness of emotions for their agents in a great variety of applications. For example Gmytrasiewicz and Lisetti (2002) quote three main reasons why emotions are essential in the design of rational agents: they can control the allocation of resources in constrained environments; they can help agents communicate their mental state to other agents in a universal vocabulary; and they must be recognized during an interaction with a human user to make it more efficient and pleasant.

Thus, the next step is to design computational models of emotions, that would translate the psychological theories exposed in the previous section into a comprehensive and computationally tractable form. The next section describes some emotional architectures.

2.3 Emotional architectures

This section describes some emotional architectures, *viz.* some computational models where emotions impact various cognitive functions.

2.3.1 Sloman

Sloman (2001) notices that great debates often arise from the confusion and indeterminacy in the very definition of complex concepts like emotions. On the contrary, he believes that architecture-based definitions of concepts are precise and allow to clearly answer questions.

Sloman thus developed the CogAff architecture of mind, a three by three grid obtained by combining two distinctions: the first one between perception, central processing and action, is introduced by Nilsson (2001) in his triple tower model; the second one is introduced by Sloman himself, between reactive, deliberative and meta-management processing levels, that appeared at different steps of evolution and provide more and more abstract and flexible processing mechanisms.

At the reactive level there are no representation, evaluation or comparison mechanisms between possible actions and their future consequences. The perception of stimuli leads to one or several competing reactions among which one is selected without any deliberation or inferences. Reactive organisms have proto-emotions (primitive versions of evolved emotions), proto-desires (needs), and a proto-mood modulating their behaviour, but they can be unable to represent or detect these states.

The deliberative level provides various deliberative abilities depending on the sophistication of the architecture: the possibility to represent, analyse, compare, evaluate and react to abstract descriptions, hypotheses or explanations. Deliberative processes interact with emotional processes, since the evaluation of abstract representations (plans, hypotheses, prospected future events...) can lead to new kinds of emotions that cannot exist in purely reactive systems. Due to these new representational abilities, emotions at the deliberative level also have a richer and more varied propositional content. Deliberative organisms modulate their behaviour depending on the context. When an organism has deliberative processes, he needs different types of interruption mechanisms to suspend them. Then he also needs to limit these interruptions, what is made possible by the addition of a filtering mechanism with a variable threshold. The new processes managing the attention threshold are called meta-management processes. They account for a human's ability to be more or less focused on what he does.

The meta-management level offers reflective meta-management processes that are crucial in a human architecture. These processes allow self-observation and self-monitoring of internal states, as well as their categorisation and evaluation, and high level mechanisms to learn and control reasoning. These new possibilities enrich the existing concepts. The third level controls processing, but it can be interrupted by the other levels of processing that can override its decisions.

Finally, Sloman believes that the architecture of an organism impacts the emotions that this organism may feel. He intends his architecture to be generic enough to describe not only emotions but also other intelligent processes like learning or even awareness.

2.3.2 Elliott

Elliott (1992) exposes in his PhD thesis in philosophy the functioning of the Affective Reasoner, a collection of LISP programs simulating the emotional behaviour of humans in an agent.

His emotion eliciting condition theory is based on Ortony and colleagues' typology. The agent has several databases: one containing his own goals, standards and preferences, and another one containing the concerns of other agents that he learned (and that thus are incomplete and possibly erroneous). The agent also has relationships with other agents, and can reason about their emotions.

An eliciting situation is appraised by a process determining the corresponding emotion eliciting conditions, that are then matched with an emotion template thanks to a database of domain-independent rules. Several emotions, even contradictory ones, can arise at the same time. Their intensity depends on some variables among a set of twenty-two possible ones (*e.g.* importance, surprisingness, tempo-

ral proximity). The agent's personality influences his appraisal of the situation, but also his expression of emotions. The agent disposes of a database containing several types of actions (*e.g.* somatic, behavioural, communicative, evaluative). Some actions are elicited among the candidate actions depending on the agent's personality and emotions, and some rules allow to resolve possible conflicts between them.

Finally the Affective Reasoner is a functional emotional program, including interesting functionalities like domain-independent appraisal, action generation, or personality management.

2.3.3 Reilly

Reilly (1996) aims at designing believable emotional agents for social interaction. Like Bates he considers the artistic aspect of this problem. He wants to create emotionally rich agents, who can feel emotions in a great variety of situations and who can express them in a great variety of ways. He thus designs *Em*, an architecture intended to help artists to design their own personalized believable emotional agents. He stresses that his aim is not to design cognitively plausible agents, even if he builds on psychologist theories. On the contrary, his framework allows to design unrealistic agents, so far as they are interesting from an artistic point of view. He thus subscribes to what Gratch and Marsella call "simulation-based approaches".

Em emotional process works as follows. First emotion generators associate the inputs with emotion structures. An emotion structure consists in a type of emotion, an intensity, an optional direction (the agent towards whom the emotion is directed), and an optional cause. Second, emotion storage functions sort the generated emotion structures into an emotion types hierarchy depending on their effects. Third an emotion combination function computes an intensity for each emotion type in the hierarchy, depending on the intensities of the emotion structures stored in this emotion type. This intensity decays over time following a decay rate specified by the user. Finally, emotions do not affect directly the agent's behaviour. They are mapped with behavioural features in Bates' sense. Then these behavioural features affect the agent's behaviour.

The artist is supposed to fill in all the user-specified elements of this architecture in order to design his own believable emotional agent. Since this task may be difficult to achieve from scratch because of the great freedom and flexibility of the system, Reilly also provides a default emotional system, *viz.* a default filling of his *Em* architecture. In this default setting the emotion generators are built on the OCC typology. Reilly thus considers his thesis as a manual to teach the artists how to design believable emotional agents with his *Em* architecture.

This architecture is integrated in an agent, and emotions interplay with several

features of the agent's behaviour, such as perception and motivation. All together this system is very complete.

2.3.4 Velàsquez

Velàasquez and colleagues (1997, 1997) designed the Cathexis architecture by grounding on notions from several fields of research including psychology and neurobiology. This architecture is set up of an emotion generation system and a behavioral system.

The emotion generation system is a network of components called *proto-specialists*, each one representing an emotion among six basic ones: anger, fear, distress/sadness, enjoyment/happiness, disgust and surprise. Actually each basic emotion is a family of related affective states sharing some characteristics like their antecedent events, expression, likely behavioral response and resulting physiological activity. Other emotions are either a variation inside a basic family or a *blend* or mixed emotion, *viz*. the simultaneous feeling of several basic emotions. Each proto-specialists has several kinds of sensors to monitor internal and external stimuli and detect the elicitation conditions of this emotion. Velàsquez grounds on Roseman's (1984) appraisal theory to describe the cognitive elicitors (*e.g*. appraisals, attributions, memory...), and he also envisages non-cognitive elicitors ranked according to Izard's (1993) view: neural, sensorimotor (*e.g*. the facial expression) and motivational (*e.g*. drives, other emotions, pain regulation). The input from these sensors either increases or decreases the intensity of the emotion. Each proto-specialist manages two thresholds of arousal: an activation threshold over which the emotion becomes active, and a saturation threshold being the maximal value of arousal for this emotion. The values of these thresholds set up the temperament of the agent. Finally each proto-specialist has a decay function controlling the duration of the emotion. Proto-specialists also manage moods, that differ from emotions by a lower activation and thus a higher duration. All proto-specialists continuously and parallely update their intensity, depending on several parameters: previous intensity, values of elicitors, and interactions (inhibitive or excitative) with other proto-specialists that are simultaneously active.

The behavioral system selects an adapted behavior depending on the current emotional state. It is set up of a network of behaviors competing for controlling the agent. The value of each behavior is computed as the sum of the values of several factors called its *releasers*, and the higher-valued behavior become active. Each behavior has two components: an expressive one determining the agent's facial expression, body posture and vocal expression; and an experiential one influencing the agent's motivations and action readiness.

The Cathexis architecture was completely implemented in a framework for the

design of emotional agents, that was then used to design Simòn, a toddler agent with which the user can interact at various levels in order to assess the underlying architecture. Finally, this is a complete and original architecture that takes into account not only the cognitive aspect of emotions but also their physiological and biological aspects.

2.3.5 Gratch and Marsella

Gratch and Marsella developed the EMA agent, endowed with a domain-independent model of emotions built on Lazarus' relational theory of emotions[2]. EMA's current mental state is represented with a complex mental structure, called Causal Interpretation, designed by the authors to unify in one single architecture / structure all the needs of an emotional agent. Indeed, they believe that none of the existing formalisms is rich enough to express the variety and complexity of emotions. They thus decided to pick parts from these different formalisms, and enriched a classical planning representation with concepts from decision theory, like probability and utility. The Causal Interpretation is set up of three causally linked parts: the causal history (the past), the current world (the present) and the task network (the future).

The appraisal process analyses the configuration of the Causal Interpretation w.r.t. several appraisal variables to trigger one or several emotions. The most intense emotion thus generated provides a coping opportunity, *viz.* it can induce a coping process, that is an attempt by the agent to adapt to his emotions and his environment. EMA disposes of several coping strategies adapted from the COPE model (Carver, Scheier, and Weintraub, 1989): planning, positive reinterpretation, acceptance, denial, mental disengagement, shift blame. The different strategies are assessed w.r.t. their coping potential, and the agent chooses and applies his preferred one. Its effect on the Causal Interpretation is mainly expressed in terms of intention dropping or modification of utility or probability values. The authors regret to have a too direct link between appraisal and coping, while psychology underlines the complexity of this link.

This agent finds applications in the Mission Rehearsal Exercise, a virtual world to teach decision-making in high-stress situations to militaries. Other work exists about emotional agents integrated into virtual worlds to favor the user's immersion. For example El Jed et al. (2004) design emotional agents that interact with the user's avatar in a virtual world for safely training firemen to decision-making in dangerous situations.

[2]EMA stands for "Emotion and Adaptation" in homage to Lazarus' book (1991).

2.3.6 Conclusion

The Affective Reasoner and the *Em* architecture are built on the OCC typology. Similarly, most of emotional agents described in the next sections also ground on this typology. But there also exists models based on other theories: on Roseman's theory (Velàsquez' Cathexis architecture discussed above), on Frijda's theory (Staller and Petta, 2001), on Lazarus' theory (Gratch and Marsella's EMA agent discussed above), on Oatley and Johnson-Laird's communicative theory (*cf.* (Meyer, 2006), that will be discussed in Section 2.6.2), or on Scherer's theory (Paiva et al., 2004).

Believable agents endowed with a computational model of emotions (we will call them "emotional agents") then find many applications in various domains: they help motivate students in pedagogical environments, they participate in a better immersion of the user in virtual worlds, for training or entertainment, ... In the next sections we describe some agents from some domains of applications. We will show that each domain has its own characteristics, but the emotion always plays a central role. We do not intend to be exhaustive but just to illustrate the usefulness of emotions for virtual agents.

2.4 Pedagogical agents

A great amount of work shows the positive impact of animated pedagogical agents on learning and motivation (Lester et al., 1997). Bates (1994) shows that emotions play an important role in these agents' believability. So researchers now try to endow their pedagogical agents with emotional intelligence. Section 2.4.1 illustrates how emotional expressiveness makes *Vincent* more motivating for students. Besides Elliott, Rickel, and Lester (1999) identify two complementary emotional reasoning abilities that are useful for such agents: expressing emotions, and understanding those of the student. Section 2.4.2 illustrate the integration of emotional expressiveness into *Steve*. Section 2.4.3 illustrates the integration of emotional responsiveness in *Herman the Bug*. Finally Section 2.4.4 briefly describes the *Mediating Agent*, using a BDI formalisation of the student to infer his emotion and choose the adapted pedagogical strategy depending on it.

This state of the art is not intended to be exhaustive but to illustrate the claim that emotional abilities make pedagogical agents better teachers, and more generally that emotions make virtual agents more intelligent.

2.4.1 Vincent

Vincent (Paiva and Machado, 1998) is a pedagogical agent composed of two main modules: the Mind Module manages his cognitive behaviour and the Body Module manages his physical behaviour. The Mind Module chooses a pedagogical strategy depending on the trainee's behaviour, and the Body Module acts consequently. Vincent first presented only four emotional reactions: impatience when time out, sadness on bad performance, friendliness on average performance, and happiness on high performance. But experiments revealed that after long interactions with him, trainees get annoyed by his monotonous behaviour and find it inconsistent because of sudden variations.

Paiva, Machado, and Martinho (1999) thus enrich Vincent with rich emotional abilities and a complete and stable personality. First they endow Vincent with a temperament associated with typical interaction sketches. Second they extend Vincent's behaviour space with an emotional dimension influencing the way he performs actions. Third they list all events and actions that may be relevant in the environment and fix the model of the user. Fourth they define Vincent's goals, that impact his actions and appraisals. Finally they define Vincent's emotional profile consisting in an emotional resistance (the intensity threshold necessary for the triggering of an emotion), an emotional memory (the duration of his emotions), and a set of emotional reactions. Vincent only appraises the trainee's success or failure to trigger prospect-based emotion of the OCC typology. The intensity of these emotions depends on a temporal proximity variable. When it increases beyond the agent's emotional resistance threshold, the agent actually "feels" the emotion.

Finally, the authors argue for the necessity of emotions in pedagogical agents who interact with students, and they propose a methodology to integrate emotions in existing pedagogical agents. Similar work has been conduced by Elliott, Rickel, and Lester (1999) who share the hypothesis that "affective reasoning will make pedagogical agents better teachers" (Elliott, Rickel, and Lester, 1999, p2). They thus try to integrate Elliott's Affective Reasoner in existing pedagogical agents: Steve (Rickel and Johnson, 1997) and Herman the Bug (Lester, Stone, and Stelling, 1999). They illustrate two different kinds of emotional reasoning: emotional responsiveness for Steve, and affective user modeling for Herman. The next sections describe this work.

2.4.2 Steve

Steve (Rickel and Johnson, 1997) is an animated agent who inhabits a 3D environment and teaches students to operate a high pressure air compressor. Steve is aware of the student's actions and of their effect on the environment. He monitors the stu-

dent's behaviour, and can answer his questions, help him, or demonstrate the task. Elliott, Rickel, and Lester (1999) want to make Steve a better teacher by enriching him with the emotional abilities offered by Elliott's Affective Reasoner. Steve would thus be able to feel emotions w.r.t. past, present or future events, actions, objects, and towards his student, according to the OCC typology.

Steve's emotions are a function of his goals and principles, but also of the student's presumed appraisal of the situation and of Steve's relationship with him. Such elements intervene in OCC fortunes-of-others emotions that make Steve not only interested in what the student does but also in what he feels (Steve seems to care about him). Since Steve's emotions are triggered by the cognitive appraisal of stimuli, he is able to explain their cause to the student, which is a great pedagogical tool. Moreover, Steve's current emotional state influences the subsequent appraisals: for example a negative emotional background facilitates the triggering of a negative emotion. The intensity of each emotion must be proportioned with the eliciting situation and is a function of twenty-two intensity variables that can be internal to one agent (personality) or common to all (domain-dependant).

This work shows that the Affective Reasoner can be integrated in a pedagogical agent to generate emotional responsiveness and personality. Another functionality of affective reasoning in such agent is the affective modeling of the user. This is illustrated by the same authors with the integration of the Affective Reasoner in Herman the Bug, detailed in the next section.

2.4.3 Herman the Bug

Herman the Bug (Lester, Stone, and Stelling, 1999) in an insect agent inhabiting the Design-a-Plant environment to teach students about botanical anatomy and physiology. He can realize various entertaining actions, and helps the students to solve problems while they graphically design customized plants that can only thrive in precise environmental conditions.

Elliott, Rickel, and Lester (1999) want to illustrate which are the benefits of affective user modeling abilities in such an agent. They state that the Affective Reasoner framework offers a reusable model of appraisal. Thus a pedagogical agent endowed with this model should be able to manage a concerns-of-others component (*cf.* Section 2.3.2 page 67) concerning the student, and to use it to understand his emotions. This model of the user (and thus also the inferences derived from it) obviously cannot be perfect but the agent disposes of five ways to update it: he can directly ask the user, use stereotypes about the user's personality, use contextual information like the interaction history, use affective stereotypes derived from statistics on all students, or if everything else fails infer how he would feel himself in the same situation. Knowing the student's principle is important to

notice when he is frustrated and to intervene but it is difficult since they differ from one student to another. Thanks to his internal representation of the student's dispositions, Herman interprets the presumed effect of events on him to infer his emotion. Various intensity variables intervene: the importance of the event for the student depending on his personality, the student's effort depending on the time he spent solving the problem, the student's mood inferred from the past successes or failures in the interaction history, and the student's arousal depending on his answering time span.

To conclude, Elliott and colleagues' work shows that emotional agents improve learning in several ways: by showing that he cares about his progress, the agent encourages the student; he shows and shares enthusiasm about the subject at hand; and his rich and funny personality attracts the student and makes him spend more time on learning.

2.4.4 Mediating Agent

Affective user modelling was also investigated by other researchers. Jaques et al. (2004) developed a pedagogical agent, the *Mediating Agent*, able to infer the student's emotions in order to adopt the better pedagogical strategy and enable the better conditions for learning. The model of the student is represented in X-BDI, an executable BDI formalism. This agent uses a domain-dependant desirability notion to trigger seven emotions from the OCC typology, associated with a qualitative intensity depending on OCC intensity variables. The agent then uses his beliefs about the student's emotion and profile, and the event at hand, to choose an appropriate affective tactic to help him. A scenario illustrates the assets of such affective user modelling during a pedagogical interaction.

2.4.5 Conclusion

Finally emotions have a positive impact for pedagogical agents and improve the students' learning experience. Emotional agents have a strong effect in motivating the student. Besides they also have an impact on interaction. Therefore more and more embodied conversational agents are now endowed with emotions. The next section describes some of these emotional conversational agents.

2.5 Conversational agents

As early noticed by Picard (1999), affective agents have a positive impact on the user and can reduce his frustration during interactions with a machine. The challenge is thus to endow interfaces with two main affective abilities: encouraging the

user to express emotions, and being able to recognize his expressed emotions in order to manage them.

The recognizing of the user's emotions was answered in various ways like recognizing the user's facial expression (*e.g.* with eyeglasses sensing his facial movements, as reported in (Picard, 1997)), vocal variations, or by monitoring various physiological signals. For example Prendinger and Ishizuka (2001) ground on Picard's work 1997 to deduce an emotion label from monitoring the user's physiological signals and gaze direction. But this problem is not in the scope of this state of the art. In this section we will rather get interested in the second problem: to make the user express his emotion. One way to achieve this goal is to design believable agents that can get the user's sympathy and trust. An important point when designing believable agents is to give them relevant multi-modal affective expressions.

Affective expressions in conversational agents can fill several functions: improve their believability, catch the user's sympathy, convey empathy, convey information or communicative intentions... The following sections explore some of the functions already addressed by agent designers.

2.5.1 Greta: an empathetic agent

In designing embodied conversational agents, it is important to endow them with believable affective facial expressions. Pelachaud and Bilvi (2003) are interested in endowing a conversational agent with non-verbal behaviour expressing his emotions during discourse to make him more believable.

Following Ortony (2003) the authors characterize emotions with two dimensions: a positive or negative valence, and time (past, present or future, indicating when the emotion eliciting stimulus occurs). These dimensions allow to differentiate not only emotions but also their typical facial expression. A belief network is used to match a situation with an emotion; then the corresponding dimensions are computed and associated with a facial expression.

The agent Greta was endowed with such an emotional system (de Rosis et al., 2003). It was then used to dialogue with users in medical applications where empathy with the user is crucial.

2.5.2 Max: a believable life-like agent

Becker, Kopp, and Wachsmuth (2004) are interested in modelling the dynamics of emotions over time in the multi-modal conversational agent Max. They integrate an emotional system into Max's architecture to improve this agent's lifelikeness and believability. Max is then used as a museum guide.

The authors differentiate between emotions, that are short-lasting and associated with an eliciting stimulus, and mood, that is long-lasting and undirected. Mood is influenced by emotions, and then impacts the triggering of new emotions. They focus on the temporal dynamics of emotions and their interaction with mood, and on the ways of communicating an identifiable emotion. They represent the communication of emotions with three dimensions: pleasure (positive or negative), arousal (level of stimulation) and dominance (level of control). They introduce a concept of boredom corresponding to a state of relatively low arousal because of a lack of stimulation. In a real-time interaction, this degree of boredom allows Max to exhibit natural proactive behaviours when the level of interaction is too low.

An emotion is triggered depending on external information. It then influences Max's facial expressions, gestures, speech, behaviour and cognitive functions. Actually emotions intervene at two levels in Max's architecture. First, discrete emotional labels, modulated by a continuous intensity, are matched with a facial expression and influence Max's reasoning. Second, continuous dimensions associated with the current emotion are used to modulate Max's observable physiological behaviour (*e.g.* the tonality of his voice or his eye blinking rate).

2.5.3 Facial expression of communicative intentions

Emotional expressions do not only create believability. They can also carry out a communicative function. Poggi and Pelachaud (2000) investigate the use of facial expressions to convey the speaker's communicative intention, *viz.* the performative of his speech acts.

Indeed, the speaker engages in dialogue with a general goal among three global types: requests, questions and informative acts. An agent who wants to communicate one of these communicative intentions may choose one or several (redundant or complementary) modalities to express it. Actually the speaker specifies his general type of goal depending on the context. He selects a specific performative that is adapted to convey his intention in this particular context. Various features of the context constrain the choice of a facial expression: the type of encounter, more or less formal, determining politeness expressions; the power relationship between sender and addressee, determining dominance or submission expressions; and the two agents' personalities influencing the expressed emotions.

The selected performative has an affective component, corresponding to an actual or potential affective state. For example a request may be accompanied by potential anger to highlight that the hearer should obey it. Therefore this affective component may be expressed at the same time as the speech act. It is expressed by an emotional expression, computed in terms of the Facial Action Coding System developed by Ekman and Friesen (1978).

This agent is thus able to determine the performative adapted to the expression of his communicative intention in a particular context of interaction, and to convey this performative through affective facial expressions.

2.6 Logical formalizations

In the previous sections we differentiated several types of work about emotions. First, some researchers design emotional architectures (*cf.* Section 2.3), simulating the influence of emotions on various cognitive functions and their interplay with mood or personality. The EMA agent is the most complete one, even integrating an account of coping strategies. These works thus match what Gratch and Marsella call simulation-based approaches (*cf.* Section 2.2.4). But these architectures are often complex, involving specific representations (*e.g.* Gratch and Marsella's Causal Interpretation).

Second, other researchers directly implement some emotions into their agents to reach a specific goal like making it more believable during a conversation with the user (Section 2.5) or having him motivating a student (Section 2.4). These approaches then rather match what Gratch and Marsella call simulation-driven approaches. They are not much interested in the internal functioning of emotions, and in their influence on the agent's cognitive functions like memory or coping. They mainly aim at making their agent more believable from the user's point of view, so they are not always very faithful to a psychological theory.

Moreover, even if some agents build on an existing emotional architecture (*e.g.* Steve and Herman the Bug integrate the Affective Reasoner architecture), most of them have their own specific emotional modules and are often domain-dependent and therefore not generic or reusable (for example the Mediating Agent uses a domain-dependent notion of desire and is thus limited to applications in a tutoring environment). Designers build on various psychological theories and use various technologies to formalize these theories, making it difficult to reuse these works.

Thus it seems that the agent community needs a generic reusable emotional model, faithful to psychology, and using a widespread technology ready to be implemented in a great variety of agents. Various formalisms have been explored to reach this objective: dynamic belief networks (Carofiglio and Rosis, 2005), fuzzy logic (El Nasr, Yen, and Ioerger, 2000), BDI logics (Meyer, 2004; Ochs et al., 2005), decision theory (Gmytrasiewicz and Lisetti, 2000), mixed with planning representations (Gratch and Marsella, 2004a). Among those we prefer formal logics, that have well-known advantages: a clear semantics allowing to disambiguate concepts, a great explanatory power of the agent's behaviour, and a formal verifiability allowing to reason about the formalized concepts. In particular BDI logics

propose to formalize an agent's mental attitudes. Now, according to Picard 1997 virtual emotions are precisely labels on the virtual agent's mental states. Even some philosophers support this view of emotions as mental attitudes (we will discuss Searle's view below). BDI logics thus seem particularly adapted to formalize emotions. Moreover they are philosophically founded (Bratman, 1987), and they and are already widely used to design agent architectures (Wooldridge, 2000), what guarantees that a model grounding on these logic would be reusable in a great amount of agents. In this section we will thus focus on the use of BDI logics to provide a generic and reusable formal model of emotions. Since the use of logic may seem at least surprising or really contradictory, we begin with exposing (in Section 2.6.1) Searle's thought that emotions are complex intentional states, thus mental attitudes like belief and desire. We then discuss some attempts to formalize an agent's emotions in a logical framework: Meyer uses his KARO logic (Section 2.6.2) and Ochs *et al*.'s use Sadek's rational interaction theory (Section 2.6.3). Since our aim is also to provide such a BDI logic formalisation of emotions, we may criticize the existing ones and conclude about what our own model is intended to improve (Section 2.6.4).

2.6.1 A philosophical view

Expressing emotions in a BDI logic may first seem rather contradictory. However, philosophy of mind has already considered emotions as mental attitudes, *viz*. the concepts that BDI logics propose to model. Indeed Searle (1983) defines *intentional states* as a kind of mental attitudes that concern a proposition. He then assumes that beliefs are intentional states whose *direction of fit* is mind to world, desires and intentions are intentional states whose *direction of fit* is world to mind, and emotions are intentional states with an empty direction of fit. So according to him emotions are particular *intentional states*, and thus they are mental attitudes just like belief and desire are.

Moreover, Searle (1983, pp. 48–51) gives some semi-formal definitions of various emotional states (*e.g*. fear, disappointment, remorse, regret, pride, shame...) expressed in terms of beliefs and desires. These definitions are quite close (although less formal) to those that we will provide in Chapter 4. Searle finally assumes that any affective state, and more generally any intentional state, can be expressed in terms of desires and beliefs.

This philosophical view supports the accuracy of the use of logic to handle emotions. Since emotions can be considered as particular mental attitudes, it seems natural to represent them with BDI logics, a framework that has been designed to formalize a rational agent's reasoning, and that is already well-tried with other (less debated) mental attitudes.

2.6.2 Meyer

Meyer (2006) describes the use of the KARO formalism (developed in (van der Hoek, van Linder, and Meyer, 1998)) to formalize some emotions and their effects on an agent's behaviour. More precisely, he focuses on the dynamic interplay between emotions and the agent's plans. He then builds on Oatley *et al.*'s communicative theory of emotions and define four of their six emotions (happiness, fear, sadness, anger) in terms of the agent's goals and plans. However, as he states himself, his aim is not to be strictly faithful to this psychological model.

> 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)

Meyer assumes that emotions are labels on particular mental attitudes. To represent them he introduces five operators: belief, desire, knowledge, action and ability. Moreover he disposes of a very expressive language to describe complex actions, like sequences or conditional actions. But the subsequent definitions of emotions are rather "task-oriented", since emotions only arise from situations relevant to the agent's intentions. Actually goals are extracted from desires (they are realizable desires not yet satisfied), and intentions are selected goals that the agent can reach (*viz.* he is able to perform an action that leads to this result). We believe that this excludes some situations where the same emotions could arise independently from any intention. For example I can be happy because the sun shines, while I can do no action to make it shine.

Meyer then proposes two axioms for each emotion: the first one expresses the emotion eliciting condition, *viz.* the particular combination of mental attitudes leading to the triggering of this emotion; the second one expresses the effect of this emotion on the agent's plans. For example an agent who is sad about the failure of a plan will deliberate to choose to revise either his plan or his goal. We believe that this link is too direct between emotion and action. Emotions may rather have a more subtle influence on the agent's behaviour (see for example the notion of *behavioural feature* used by Reilly, Section 2.3.3).

Finally Meyer proposes an expressive formal framework to model emotions. He accounts for their triggering as well as their effects on the agent's plans. Actually the triggering of emotions only depends on the agent's plans, thus not covering a number of other emotional situations that are not relevant to Meyer's aim. Moreover the triggering of emotions is also limited to individual aspects and does not take social standards into account; besides only a few emotions are described. Their effect is formalized by an axiom expressing how they influence the agent's

deliberation process, grounding on action tendencies in the sense of Frijda (1986). Thus in our sense this effect is fixed, *viz.* each emotion always has the same kind of effect on the agent's behaviour. On the contrary in this work we will formalize the impact of emotions on behaviour in terms of *coping* strategies in the sense of Lazarus and Folkman (1984) (*cf.* Chapter 8). This way, each emotion can lead to the use of any strategy depending on the context, what allows a more varied behaviour making the agents more believable. Yet Meyer's aim is not to design believable agents but efficacious ones, using emotions as heuristics for decision-making: his formalization of action tendencies may thus be more adapted for his aim. Moreover this formalization has well-established semantics and axiomatics and allows to disambiguate emotions; in this sense the simplifications mentioned above were necessary. Finally Meyer has also designed a programming language to actually integrate such emotions into agents (Dastani and Meyer, 2006).

2.6.3 Ochs *et al.*

Ochs et al. (2005) give a formalization of emotions based on Sadek's Rational Interaction Theory (1992), a logic of belief, intention, and uncertainty. They then define abbreviations for particular mental attitudes called present and future (un)desirability in terms of the (actual or expected) realization of the agent's choices. They next define four emotions from the OCC typology (joy, sadness, hope, and fear) that actually match these four mental attitudes. As in Meyer's work, the authors focus on the individual aspects of the defined emotions. Indeed their aim is to compute the facial expression of an animated agent. They also discuss how such an agent can express several emotions at the same time through a mixed facial expression.

There seem to be some formal problems in the definitions exposed. First these mental attitudes do not involve beliefs but uncertainty (that are exclusive from beliefs, *viz.* if an agent is uncertain about a proposition then he does not believe it to be true or false). Thus an agent can feel an emotion about a proposition only if he is uncertain about this proposition. If an event makes me believe that a given proposition is true, I am no more uncertain and thus I feel no emotion. For example I cannot be happy that the sun is shining unless I have doubts that it is.

Moreover Sadek's choice is supposed to be strongly realistic, *viz.* when an agents believes a proposition to be true, he also chooses it to be true (*cf.* Chapter 3). Then any event that adds a belief to the agent's knowledge base also adds the corresponding choice. So the emotions cannot either be defined as a belief that a choice was realized, or the agent would feel joy whatever the occurred event. We believe that actually emotions do not arise from the realization of goals but from the one of desires. Indeed desires are not strongly realistic.

Finally this work is another attempt to disambiguate emotions through logical definitions. The focus is not the same as in Meyer's work, so the result is less formal. However it also supports the use of logical representation of emotions, considered as mental attitudes. Moreover it explores some other aspects like emotional blending. This work was recently enriched to account for empathetic emotions (happy for, sorry for, ...) and their role in human-computer interaction (*cf*. (Ochs, Sadek, and Pelachaud, 2007) or (Ochs, Pélachaud, and Sadek, 2006)). Once again the focus is not on providing a sound and complete axiomatization of emotions (indeed no semantics is given for the modal operator representing emotions), but rather on the disambiguation of the effects of emotions.

2.6.4 Conclusion

These two logical formalizations adopt two different views on the use of BDI logics. Meyer focuses on providing a well-established semantics and axiomatics of emotions, that allow to reason about their effects on the agent's plans. On the contrary Ochs *et al*. rather use logical formulas to disambiguate their definitions of emotions and then focus on their expression or effect on the interaction. Nevertheless these two works show that it is possible to provide a logical description of such a complex phenomenon as emotions. Such a description is inevitably simplistic but it allows to disambiguate the phenomenon and to integrate emotions into virtual agents. Indeed a great number of agents already build on BDI architectures (Wooldridge, 2000).

However we would like to highlight a number of limitations of these works. First, both formalisations focus on the individual aspects of emotions and neglect the social aspects, in particular the influence of social norm on the construction of emotions. However this influence in essential, all the more in studying the influence of emotions on interaction. Second, they define rather few emotions, and are not always very faithful to the original psychological definitions on which they build. Nevertheless we believe that designers of believable agents should trust the psychologists who have tailored their definitions for decades. Finally, they both neglect the concept of desire: we can say that they are too task-oriented since they only consider goals, choices or plans. On the contrary we believe that desires are essential in the triggering of human emotions. Their definitions are thus not generic enough to capture all the variety of situations that can trigger emotions.

In the light of this review of existing work, we can now define more precisely the aim of this thesis.

2.7 Conclusion

This chapter has proven that emotions are worth being integrated into virtual agents, and improve their behaviour in a variety of applications, ranging from pedagogical agents to virtual worlds. We then noticed that the great amount of works to design such emotional agents makes it necessary to design a generic model of emotions that designers could reuse in various agents and for various applications. We thus discussed several existing attempts to provide such a formal model, in particular by using BDI logics that are a widespread technology to describe rational agents architectures. Philosophy of mind supports the underlying hypothesis that emotions can be considered as a kind of mental attitudes. We then criticized several drawbacks of these existing BDI formalisations of emotions, in particular their lack of genericity due to the use of goals instead of desires, and their focus on individual aspects at the expense of social ones. Moreover we can regret that these approaches do not take advantage of the power of BDI logics to reason about emotions and prove some properties about them.

On the contrary our aim in this thesis, as stated in the introduction, is to propose a formal model of emotions. Following Searle and Meyer, we consider emotions as particular mental states and choose to build our model on a BDI logic. This model should meet several requirements. **First** it must be as faithful as possible to the OCC typology. Indeed psychology has already experimented its definitions for decades, so we have to trust them. Moreover our aim is not to design efficacious agents optimized for a given application, but rather to propose a model as generic as possible, independent from any domain of application. We thus stay close to the underlying psychological theory and let each agent designer optimize our model for his own application if needed. **Second** we want to formalize a great number of different emotions because diversified emotional expressions are essential to make agents more believable. Actually we will propose definitions for twenty emotions of the OCC typology, *viz.* all but the object-based ones. Our model will thus offer the richest set of formal definitions of emotions up to date. We then let the agent designers choose which emotions among the formalized ones they need for their particular application. **Third** we want to consider the social aspects of emotions and we will then introduce a modal operator accounting for social standards. **Fourth** we want to stay as generic as possible, in particular our model should not be task-oriented as it is the case for Meyer and Ochs *et al.*'s accounts. We will thus express emotions in terms of desires rather than in terms of goals (this choice will be further discussed along with our other formalization choices in Chapter 4). **Fifth** we want to be able to reason about emotions and prove some of their properties. We will then strive to keep our logic not too complex, in order to be able to prove its soundness and completeness. The proofs of a set of theorems

about emotions constitute the main originality of our work.

The next part of this thesis is dedicated to our formalization of the OCC typology. We begin with introducing our particular BDI framework (Chapter 3), then proceed with providing and justifying our formal definitions of twenty emotions from the OCC typology (Chapter 4), and finally expose and prove various properties of these emotions (Chapter 5). This work sets up the core of this thesis.

Part II

Logical formalization of emotions

Feelings are not supposed to be logical. Dangerous is the man who has rationalized his emotions

David Borenstein

Chapter 3

Logical framework

Logic: The art of thinking and reasoning
in strict accordance with the limitations and
incapacities of the human misunderstanding.
Ambrose Bierce

3.1 Introduction

As we showed in the state of the art, the agent community recently got very interested in emotional agents. Unfortunately all the approaches design their own emotional model starting from scratch, *viz*. from one of the numerous existing psychological theories. Thus, there exists a great variety of computational models of emotions, depending on the application, the context of use, or the underlying formalism (*cf*. Chapter 2). We believe that the agent community needs a generic formal model of emotions, and we assume that BDI logics (*viz*. logics of *Belief*, *Desire* and *Intention*, *e.g*. Cohen and Levesque (1990), Rao and Georgeff (1991; 1992), Sadek (1992), Herzig and Longin (2004)) allow to develop such a model.

Indeed emotions are complex phenomena, their psychological definition is often abstract, ambiguous and non consensual, what leads to many debates (*cf*. Chapter 1). For example two of the most cited psychological theories disagree on the definition of hope: according to Lazarus (1991) hope is the emotional state arising when expecting something bad to happen but still envisaging something better; according to Ortony, Clore, and Collins (1988) the probabilities are inverted and hope arises when you expect something good to happen but you are not sure of it. Psychological definitions of emotions are thus often subjective, whereas emotional agent designers need clear definitions, ready to be implemented in their agents. Formal logic provides such a universal vocabulary, with a clear semantics. It also

allows reasoning, planning and explanation of an agent's behaviour. The logical formalization of a phenomenon can even reveal problems that do not appear intuitively. BDI logics (Cohen and Levesque (1990), Rao and Georgeff (1991,1992), Sadek (1992), Herzig and Longin (2004), Wooldridge (2000)), that ground on the philosophy of language, mind, and action (*cf.* Bratman (1987), Searle (1969, 1983)), propose to model agents *via* some key concepts such as action and *mental attitudes* (beliefs, goals, intentions, choices...). This framework is commonly used in the agent community and offers well-known interesting properties: great explanatory power, formal verifiability, rigorous and well-established theoretical frame (from the point of view of both philosophy and formal logic).

In this chapter we thus provide a BDI framework to model emotions. Our logic is a propositional modal logic. While there is a large consensus about the logic of belief, several concepts of desire exist in the literature. Broadly these concepts fall in two categories. In the first one desire is viewed as something 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.) In the second category desires rather correspond to general preferences whose existence does not depend on its satisfaction, such as an agent's general preference of sunny days over rainy days. In many BDI approaches the second option is taken, and moreover desires are strongly connected to beliefs (*cf.* the goal concept of Cohen and Levesque (1990) or Rao and Georgeff (1991), or the choice concept of Sadek (1992) or Herzig and Longin (2004)). Here we opt for the second view, but we do not connect desires to beliefs as the other approaches do. Finally, it turns out that the intention component can be omitted in our framework, as well as any "intermediate" mental attitude (choice, achievement goal, persistent goal, intention...), since they are not needed to describe the OCC emotions (but note that intentions will be central when it comes to coping strategies (*cf.* Chapter 8). Moreover, our framework also uses time, action, and ideality operators. Our notion of ideality allows to represent an agent's internalized standards, be they moral, legal...

This chapter explains this particular BDI framework in details. Section 3.2 exposes the semantics of our modal operators. Section 3.3 describes their axiomatics. Section 3.4 proves the soundness and completeness of our logic.

3.2 Semantics

Let $AGT = \{i, j, k, ...\}$ be the set of agents and $ACT = \{\alpha, \beta, \gamma, ...\}$ the set of actions.

3.2.1 Kripke models

We use a standard possible worlds semantics, and a model $\mathcal{M}$ is a triple $\langle W, V, \mathscr{R}\rangle$ where W is a set of possible worlds, V is a truth assignment which associates each world w with the set V_w of atomic propositions true in w, and $\mathscr{R} = \{\mathscr{A}, \mathscr{B}, \mathscr{P}, \mathscr{D}, \mathscr{G}, \mathscr{I}\}$ is a tuple of mappings made up of the following:

- $\mathscr{A} : ACT \to (W \to 2^W)$ associates each action $\alpha \in ACT$ and possible world $w \in W$ with the set $\mathscr{A}_\alpha(w)$ of possible worlds resulting from the performance of action α in w. We impose that for every $w \in W$: (1) if $w' \in \mathscr{A}_\alpha(w)$ and $w'' \in \mathscr{A}_\beta(w)$ then $w' = w''$; (2) if $w \in \mathscr{A}_\alpha(w')$ and $w \in \mathscr{A}_\beta(w'')$ then $w' = w''$. This imposes that actions are organized into histories[1] and take one time step. Therefore, actions are deterministic in the future and in the past;

- $\mathscr{B} : AGT \to (W \to 2^W)$ 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. All these accessibility relations are serial, transitive and euclidian;

- $\mathscr{P} : AGT \to (W \to 2^{2^W})$ associates each agent $i \in AGT$ and possible world $w \in W$ with a set of sets of possible worlds $\mathscr{P}_i(w)$. According to Chellas (1980, chap. 8), these sets of possible worlds are called *neighbourhoods* of w. We will impose below that neighbourhoods are "big" subsets of $\mathscr{B}_i(w)$: every $U \in \mathscr{P}_i(w)$ intuitively contains more elements than its complement $\mathscr{B}_i(w) \setminus U$. The following constraint is slightly weaker[2] and does not completely capture this intuition: for every $w \in W$, if $U_1, U_2 \in \mathscr{P}_i(w)$ then $U_1 \cap U_2 \neq \emptyset$. In other words, if φ is probable (*viz.* φ is true in all the worlds of neighbourhood), then $\neg\varphi$ is not (since each other neighbourhood contains at least one world where φ is true). Finally, in order to ensure that at least tautologies are probable, we impose that $\mathscr{P}_i(w) \neq \emptyset$ for every $w \in W$;

- $\mathscr{D} : AGT \to (W \to 2^W)$ 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. All these accessibility relations $\mathscr{D}_i$ are serial;

- $\mathscr{G} : W \to 2^W$ associates each possible world $w \in W$ with the set $\mathscr{G}(w)$ of possible worlds in the future of w. This accessibility relation is a linear order (reflexive, transitive, antisymmetric and total).

[1] It does not impede the parallel execution of several actions, but it guarantees that all these parallel actions lead to the same world (at the same time in the same history).

[2] There are neighbourhoods satisfying our constraints while "gathering" less than 50 % of the worlds, *cf.* Walley and Fine (1979).

- $\mathscr{I} : AGT \rightarrow (W \rightarrow 2^W)$ associates each agent $i \in AGT$ and possible world $w \in W$ with the set $\mathscr{I}_i(w)$ of worlds ideal for agent i. In these ideal worlds all the (social, legal, moral...) obligations, norms, standards... that the agent i has internalized (accepted for himself) hold. All these relations are serial.

Moreover, we impose some constraints involving two or more accessibility relation types:

- if $w \in \mathscr{B}_i(w')$ then $\mathscr{P}_i(w) = \mathscr{P}_i(w')$ and $\mathscr{D}_i(w) = \mathscr{D}_i(w')$, which ensures that agents are aware of their probabilities and desires[3];
- if $(\mathscr{B}_i \circ \mathscr{A}_\alpha)(w) \neq \emptyset$ then $(\mathscr{A}_\alpha \circ \mathscr{B}_i)(w) \subseteq (\mathscr{B}_i \circ \mathscr{A}_\alpha)(w)$, which ensures that agents do not forget their previous alternatives ("no forgetting", alias "perfect recall" (Fagin et al., 1995)). This is based on the hypothesis that actions are public, *viz.* that they are perceived correctly and completely by every agent;
- $U \subseteq \mathscr{B}_i(w)$ for each $U \in \mathscr{P}_i(w)$, which entails that belief implies probability;
- $\mathscr{G} \supseteq \mathscr{A}_\alpha$ for each α, which ensures that the future of every w contains the worlds resulting from the performance of actions in w;
- finally, for the sake of simplicity, we make the hypothesis that what the agent desires persist: if $w\mathscr{G}w'$ then $\mathscr{D}_i(w) = \mathscr{D}_i(w')$;
- we make the same hypothesis for (social, legal, moral...) obligations, norms, standards... that the agents have accepted for themselves: if $w\mathscr{G}w'$ then $\mathscr{I}_i(w) = \mathscr{I}_i(w')$.

We are aware that in general these last two constraints are too strong, but they are quite realistic for rather short time intervals like a small dialog.

3.2.2 Modal operators and language

We associate modal operators to these mappings:

- $After_\alpha\, \varphi$ reads "φ is true after performance of action α";
- $Before_\alpha\, \varphi$ reads "φ is true before performance of action α";

[3]Due to the transitivity and euclidianity of the $\mathscr{B}_i$ relations, agents are also aware of their beliefs, *viz.* we can derive the following property: if $w \in \mathscr{B}_i(w')$ then $\mathscr{B}_i(w) = \mathscr{B}_i(w')$.

- $Bel_i\,\varphi$ reads "agent i believes that φ";
- $Prob_i\,\varphi$ reads "for i φ is more probable than $\neg\varphi$";
- $G\varphi$ reads "henceforth φ is true";
- $H\varphi$ reads "φ has always been true in the past";
- $Idl_i\,\varphi$ reads "ideally it is the case for i that φ".
- $Des_i\,\varphi$ reads "φ is desirable for i";

$ATM = \{p, q...\}$ is the set of atomic formulas. Every atomic formula is a complex formula. If φ and ψ are two complex formulas and $\Box$ is one of the above modal operators, then $\neg\varphi$, $\varphi \vee \psi$ and $\Box\varphi$ are complex formulas. The set of complex formulas is noted $FORM = \{\varphi, \psi...\}$.

As usual, if φ and ψ are complex formulas, we define $\varphi \wedge \psi$ and $\varphi \rightarrow \psi$ as abbreviations of complex formulas. We also define the following abbreviations:

- $Happens_\alpha\,\varphi \stackrel{def}{=} \neg After_\alpha\,\neg\varphi$ reads "α is about to be performed, after which φ";
- $Done_\alpha\,\varphi \stackrel{def}{=} \neg Before_\alpha\ \neg\varphi$ reads "α has just been performed, and φ was true before";
- $F\varphi \stackrel{def}{=} \neg G\neg\varphi$ reads "φ is true or will be true at some future instant";
- $P\varphi \stackrel{def}{=} \neg H\neg\varphi$ reads "φ is or was true".

3.2.3 Truth conditions

The truth conditions are standard for almost all of our operators:

$$w \Vdash \Box\varphi \quad \text{iff} \quad w' \Vdash \varphi \text{ for every } w' \in \mathscr{R}_\Box(w)$$

where $(\Box, \mathscr{R}_\Box)$ is either $(After_\alpha, \mathscr{A}_\alpha)$ with $\alpha \in ACT$, or $(Bel_i, \mathscr{B}_i)$ with $i \in AGT$, or $(Des_i, \mathscr{D}_i)$ with $i \in AGT$, or $(G, \mathscr{G})$, or $(Idl_i, \mathscr{I}_i)$ with $i \in AGT$.
For the dual operators we have:

$$w \Vdash \boxminus\varphi \quad \text{iff} \quad w' \Vdash \varphi \text{ for every } w' \text{ so that } w \in \mathscr{R}_\boxminus(w')$$

where $(\boxminus, \mathscr{R}_\boxminus)$ is either $(Before_\alpha, \mathscr{A}_\alpha)$ with $\alpha \in ACT$, or $(H, \mathscr{G})$.

The truth conditions illustrate that the operator H is interpreted as the dual of G and $Before_\alpha$ is interpreted as the dual of $After_\alpha$. Moreover:

$$w \Vdash Prob_i\, \varphi \quad \text{iff} \quad \text{there exists } U \in \mathscr{P}_i(w) \text{ so that for every } w' \in U,\ w' \Vdash \varphi.$$

Hence φ is probable for i in w if there is a "big" subset of $\mathscr{B}_i(w)$ where φ holds.

Validity of a formula φ in the class of all Kripke models obeying our semantic constraints is defined as usual, and is noted $\models \varphi$.

3.3 Axiomatics

We now introduce a set of axioms that our modal operators have to satisfy. First we recall what a normal modal operator is.

3.3.1 Normal operators

$\Box$ is a normal operator iff the axiom (K-$\Box$) and the necessitation rule (RN-$\Box$) are valid. This characterizes a possible worlds semantics in the sense of Kripke (1963). (RN-$\Box$) means that if φ is true then it is necessarily true.

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

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

We recall that in any normal modal logic, the semantics validates in particular[4] the factorisation rule (C-$\Box$) and the inference rule (RM-$\Box$) expressing that the set of necessary formulas is closed under implication:

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

$$(\Box A \wedge \Box B) \to \Box(A \wedge B) \qquad \text{(C-}\Box\text{)}$$

The dual of $\Box$ is denoted $\Diamond$ and obeys the following theorem, meaning that if A is true in every world and there exists a world where B is true then there exists a world where A and B are both true:

$$(\Box A \wedge \Diamond B) \to \Diamond(A \wedge B) \tag{3.1}$$

$\Diamond$ also obeys the following inference rule (Chellas, 1980, Theorem 4.4 p.116):

$$\frac{A \to B}{\Diamond A \to \Diamond B} \qquad \text{(RK-}\Diamond\text{)}$$

For more details on the formal properties of normal modal logics, *cf.* (Chellas, 1980, chap. 4).

[4] We only recall the principles that will be needed to prove some theorems in Chapter 5.

3.3.2 Action

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

$$Happens_\alpha\,\varphi \rightarrow After_\beta\,\varphi \qquad \text{(CD-HA)}$$
$$Done_\alpha\,\varphi \rightarrow Before_\beta\,\varphi \qquad \text{(CD-DB)}$$
$$\varphi \rightarrow After_\alpha\,Done_\alpha\,\varphi \qquad \text{(CONV-AD)}$$
$$\varphi \rightarrow Before_\alpha\,Happens_\alpha\,\varphi \qquad \text{(CONV-BH)}$$

(CD-HA) and (CD-DB) are the axioms of common determinism, and entail that actions 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.

In the following, the notation $i{:}\alpha$ reads "agent i is the author of action α".

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:

$$Bel_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \qquad \text{(D-}Bel_i\text{)}$$
$$Bel_i\,\varphi \rightarrow Bel_i\,Bel_i\,\varphi \qquad \text{(4-}Bel_i\text{)}$$
$$\neg Bel_i\,\varphi \rightarrow Bel_i\,\neg Bel_i\,\varphi \qquad \text{(5-}Bel_i\text{)}$$

Thereby an agent's beliefs are consistent (D-Bel_i), and an agent is aware of what he believes (4-Bel_i) and of what he does not believe (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:

$$G\varphi \to \varphi \qquad \text{(T-}G\text{)}$$
$$(F\varphi \wedge F\psi) \to F(\varphi \wedge F\psi) \vee F(\psi \wedge F\varphi) \qquad \text{(3-}F\text{)}$$
$$G\varphi \to GG\varphi \qquad \text{(4-}G\text{)}$$
$$H\varphi \to \varphi \qquad \text{(T-}H\text{)}$$
$$(P\varphi \wedge P\psi) \to P(\varphi \wedge P\psi) \vee P(\psi \wedge P\varphi) \qquad \text{(3-}P\text{)}$$
$$H\varphi \to HH\varphi \qquad \text{(4-}H\text{)}$$
$$\varphi \to GP\varphi \qquad \text{(CONV-GP)}$$
$$\varphi \to HF\varphi \qquad \text{(CONV-HF)}$$

(T-G) and (T-H) mean that future and past are taken in a broad sense: if a proposition is always true in the future or in the past, it is true in particular in the present.

(4-G) et (4-H) express the transitivity of time in past and future: if φ is true in all the futures (resp. in all the pasts), then φ is also true in all the futures of these futures (resp. in all the pasts of these pasts).

(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 the time is linear in the future and in the past. Figure 3.1 illustrates this on a model where $w_0 \Vdash Bel_i\, G\varphi$, $w_1 \Vdash \neg Bel_i\, \neg(\neg\varphi \wedge F\varphi)$, $w_1 \Vdash \neg Bel_i\, \neg G\varphi$ and $w_1 \Vdash \neg Bel_i\, \neg G\neg\varphi$.

(CONV-GP) and (CONV-HF) are the conversion axioms.

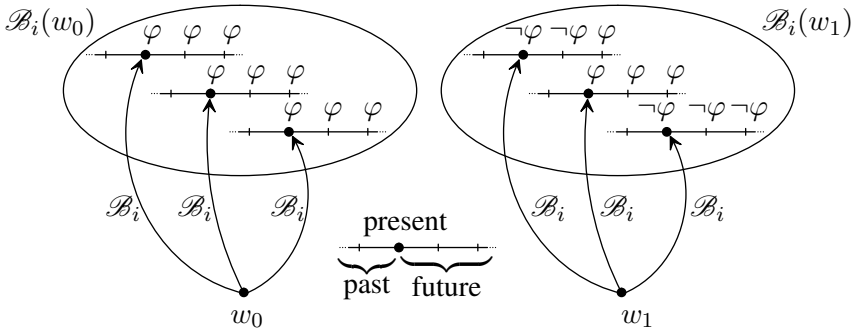

Figure 3.1: Representation of linear time

One might object that at least future should be branching. For us, what is

important is not the nature of time but rather the perception that agents have of it: we choose to represent the diversity of futures through different histories whose present are different epistemic worlds (*cf*. Figure 3.1). Thus, although we represent time in a linear setting, the agent has a branching perception of it.

3.3.5 Probability

The probability operators correspond to the notion of weak belief, based on the notion of subjective probability measure (this aspect is captured semantically by the fact that probable worlds belong to the set of believed worlds).

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 logic (Chellas, 1980, Theorem 4.3).

The semantical conditions validate the following principles:

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

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

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

Thereby an agent's probabilities are consistent (D-$Prob_i$) and closed under implication (RM-$Prob_i$). Moreover tautologies are probable (RN-$Prob_i$).

3.3.6 Desirability

Its logic is identical to the one of the standard deontic logic (SDL) and is also expressed in terms of ideal worlds: the logic associated with the operators Des_i is **KD**, *viz*. 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 desires consistent.

An agent's desires are individual preferences. They can be unrealistic because we do not impose that $\mathscr{B}_i(w) \cap \mathscr{D}_i(w) \neq \emptyset$: an agent can desire to be in various states that he currently believes to be impossible.

We stress that in principle (*e.g*. Lang, Van Der Torre, and Weydert (2002), Castelfranchi and Paglieri (2007)), desires are closed neither under implication nor under conjunction: I can desire to marry Ann and desire to marry Beth without desiring to be a bigamist. However, for the sake of simplicity, our Des_i operators are normal and thus do not respect these principles (*viz*. they are closed under both conjunction and implication).

3.3.7 Ideality

The notion of ideality considered here is an obligation that is taken in a large sense: it embraces all the rules imposed on the agent by some external authority, provided that the agent has internalized them, *viz.* accepted them for himself. They can be explicit (like laws) or more or less implicit (like social or moral obligations). Therefore they can be said to correspond to the agent's moral values. They are a kind of social preferences stemming from the groups to which the agent belongs, and thus differ from the agent's personal desires that are expressed by means of the Des_i operator.

The logic of ideality is the Standard Deontic Logic (Jones and Carmo, 2002), *viz.* the normal modal logic **K** plus the following axiom making ideals consistent:

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

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 desires:

$$Prob_i\,\varphi \rightarrow Bel_i\,Prob_i\,\varphi \qquad \text{(4-MIX1)}$$
$$\neg Prob_i\,\varphi \rightarrow Bel_i\,\neg Prob_i\,\varphi \qquad \text{(5-MIX1)}$$
$$Des_i\,\varphi \rightarrow Bel_i\,Des_i\,\varphi \qquad \text{(4-MIX2)}$$
$$\neg Des_i\,\varphi \rightarrow Bel_i\,\neg Des_i\,\varphi \qquad \text{(5-MIX2)}$$

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

Then the following axioms express that actions are public:

$$Done_\alpha\,\top \rightarrow Bel_i\,Done_\alpha\,\top \qquad \text{(4-MIX3)}$$
$$\neg Done_\alpha\,\top \rightarrow Bel_i\,\neg Done_\alpha\,\top \qquad \text{(5-MIX3)}$$

From these axioms plus (D-Bel_i), we can easily prove their converse, and we thus have equivalences.

We express the inclusion of probable worlds in the set of epistemic worlds through the following axiom:

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

which allows to derive the following intuitive theorems:

$$Bel_i\,\varphi \rightarrow Prob_i\,\varphi \tag{3.2}$$
$$Prob_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \tag{3.3}$$

(3.2) reads "if agent i believes φ then for him φ is more probable than $\neg\varphi$". (3.3) reads "if agent i considers that φ is probable then he envisages that φ can be true".

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

$$G\varphi \rightarrow After_\alpha\,\varphi \tag{GA-MIX}$$
$$H\varphi \rightarrow Before_\alpha\,\varphi \tag{HB-MIX}$$

Finally, desires and undesires persist (*viz.* they are preserved through time):

$$Des_i\,\varphi \rightarrow GDes_i\,\varphi \tag{Pers-Des_i}$$
$$\neg Des_i\,\varphi \rightarrow G\neg Des_i\,\varphi \tag{Pers-$\neg Des_i$}$$

These two principles entail that we have an equivalence.

For the same reasons, the ideals imposed to an agent also persist:

$$Idl_i\,\varphi \rightarrow GIdl_i\,\varphi \tag{Pers-Idl_i}$$
$$\neg Idl_i\,\varphi \rightarrow G\neg Idl_i\,\varphi \tag{Pers-$\neg Idl_i$}$$

These two principles entail that we have an equivalence.

We stress that we have **not** supposed that the agents are aware of imposed ideals. Indeed, intuitively, the principles $Idl_i\,\varphi \rightarrow Bel_i\,Idl_i\,\varphi$ and $\neg Idl_i\,\varphi \rightarrow Bel_i\,\neg Idl_i\,\varphi$ would be too strong, and they are not valid in our semantics.

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

$$Bel_i\,After_\alpha\,\varphi \wedge \neg Bel_i\,After_\alpha\,\bot \rightarrow After_\alpha\,Bel_i\,\varphi \tag{NF-Bel_i}$$

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 α is not executable"). Otherwise, if $Bel_i\,After_\alpha\,\top$ 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).

3.4 Soundness and completeness

We call $\mathcal{L}$ the logic thus axiomatized, and write $\vdash_{\mathcal{L}} \varphi$ iff φ is a theorem of $\mathcal{L}$.

Theorem (Soundness and completeness). $\vdash \varphi \quad \textit{iff} \quad \models \varphi$.

Sketch of proof. *It is a routine task to check that all the axioms correspond to their semantic counterparts. It is routine, too, to check that all of our axioms are in the Sahlqvist class, for which a general completeness result exists (Sahlqvist (1975), Blackburn, de Rijke, and Venema (2001))* □

3.5 Conclusion

We thus dispose of a set of modal operators to describe the agents' mental attitudes and reasoning abilities. Our logic is a propositional modal logic. An agent has beliefs, probabilities, personal desires and internalized social ideals, and he can reason about time and action.

The next step of our work now consists in characterizing emotions in terms of these concepts. The soundness and completeness will then be important to reason properly about these emotions and prove some of their properties.

Chapter 4

Formal definitions of emotions

Life obey no logic,
why do we want to deduce its meaning with logic?
Gao Xingjian

4.1 Introduction

In this chapter we tackle the core of this work: formalizing emotions in the logical framework exposed in the previous chapter. We have shown in Chapter 1 that cognitive appraisal theories are best adapted to reason about emotions and their cognitive antecedents, *viz*. the particular mental state that cause them. Among the various available cognitive appraisal theories, we have already discussed in Section 1.4 (page 57) why we choose to formalize the so-called OCC typology due to Ortony, Clore and Collins (1988, *cf*. Section 1.3.3). Indeed it is easier to be understood by a computer scientist than the other theories, since it was intended for Artificial Intelligence. It has thus become the most cited approach in the agent community. This gives it a kind of legitimacy, since it seems to fit the requirements for most of existing emotional agents. Of course this argument does not ensure that this theory is better than the other ones, so this problem will be discussed again in Chapter 7 when using our model to assess the underlying theory.

In this chapter we start from the OCC typology and try to formalize its definitions in our formal language presented in Chapter 3. Actually we only formalize two branches of the OCC typology. The event-based branch of the OCC typology contains emotion types whose eliciting conditions depend on the evaluation of an event w.r.t. the agent's goals. *Desirability* is a central intensity variable accounting for the impact that an event has on an agent's goals, *viz*. how it helps or impedes their achievement. The agent-based branch of the OCC typology contains emotion

types whose eliciting conditions depend on the judgement of the praiseworthiness of an action, with respect to standards. 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.

There are some difficulties in such an initiative. The first one is to understand what is meant in the definition: the concepts involved can be ambiguous, and we have to interpret the definition before formalizing it. The second one is to find an adapted formalization for these concepts: indeed a logical language is inevitably far less expressive that natural language, and it is difficult to match complex psychological concepts with a limited set of modal operators. Therefore, our formal definitions depend on our interpretation of Ortony *et al.*'s informal definitions, and this interpretation is subject to debate. That's why in the following sections we will not only expose our formalizations but also discuss our choices for each one of them. Moreover we will also support the accuracy of our choices by showing that our definitions can capture the situations that Ortony *et al.* use in their book to illustrate their emotion types.

In the following we present ten pairs of opposite emotions (that can be entire groups or subgroups in the OCC typology), while always respecting the same structure:

- we begin with giving Ortony and colleagues' informal definitions of one or several close pairs of emotions; we immediately propose a formal definitions of these emotions and apply them to Ortony *et al.*'s examples (the page numbers always refer to their book (Ortony, Clore, and Collins, 1988));
- we then discuss our choices of formalization and support them by giving some examples;
- finally we compare Ortony *et al.*'s definitions of these emotions with the definitions of close emotions from Lazarus' theory (we always refer here to his book (Lazarus, 1991), presented in details in Section 1.3.2).

4.2 Well-being emotions

The emotion types in this group have eliciting conditions focused on the desirability for the self of an event.

4.2.1 Well-being emotions

Definition by OCC (Joy and distress). *An agent feels joy (resp. distress) when he is pleased (resp. displeased) about a desirable (resp. undesirable) event.*

Our formal definition 1 (Joy and distress).

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

Example by OCC. *For example in (Ortony, Clore, and Collins, 1988, p. 88), when a man i hears that he inherits of a small amount of money from a remote and unknown relative k ($Bel_i\,(m \wedge d)$), he feels **joy** because he focuses on the desirable event ($Des_i\,m$). On the contrary, this man does not feel distress about his relative's death, because since he did not know him we can guess that 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 ($Des_j\,\neg o$).*

4.2.2 Choices of formalization

4.2.2.1 Desirability

In this definition (and in all the following ones), we represent something "desirable" for the agent i through the modal operator Des_i . Other researchers who tried to account for emotions in a logical framework have rather characterized this desirability through the achievement of a goal (Ochs et al., 2005; Meyer, 2004). So why do we use desires rather than goals? Let's illustrate this choice with two examples.

First, we suppose that a man has to go to the dentist to treat caries. He believes that it will hurt so he desires not to go, he does not like to go. But he also knows that it would be worse if he does not go, so he decides to and adopt the goal or intention to go. The day of the operation he is probably afraid about it. Now, we suppose that the dentist calls him to cancel the operation. The man would probably feel relieved about this, although his intention has just failed. What makes him happy here is his satisfied desire (not to go to the dentist) and not his goal.

Second we suppose that a woman hears that some singer that she likes has just died in an accident. She feels quite sad about this because she liked this person. However we cannot say that she had the intention that the singer does not die, since this does not depend on her, she cannot do anything to prevent it. Actually this woman does not intend that the singer stays alive, but she desires it.

Finally, desires are something weaker than intentions (to desire something does not imply that it is possible to make it happen). When they are realistic some desires can be selected and become goals or intentions, but many times desires are contrary to intentions. For example I intend to go to work every morning while I desire to stay at home. Desires are closer to what Lazarus calls ego-involvement,

they represent the agent's very personality, what makes him really different from the other agents. That's why in the following we always use desires rather than goals or intentions to represent what makes an agent feel emotions. Thus we formalize *desirability* through our Des_i operators.

4.2.2.2 Belief or surprise

Some researchers suppose that the agent needs to be surprised in order to feel joy. We agree that surprise can increase the intensity of joy: indeed according to Ortony and colleagues the intensity of event-based emotions is impacted by the likelihood of the event; and many researchers consider the novelty (or surprisingness) of the stimulus as an appraisal variable. But we believe that surprise is not necessary to feel joy: for example a couple who has a baby has awaited him for nine months, they are thus not surprised at all the day of his birth, and nevertheless they feel joy. We thus do not impose in the definition that the event at hand is surprising for the agent.

4.2.3 Discussion

According to Lazarus, the *core relational theme*[1] for joy is "reasonable progress towards the realization of our goals". But he also specifies that happiness (or joy) happens when "we have gained or are gaining what we desire". This seems to support our choice of using desires rather than goals in our definitions.

Moreover Lazarus imposes that the general existential background should be favorable, since the presence of situations inducing negative emotions can at the same time decrease or even prevent joy. This is not imposed by Ortony *et al.*: in their view the individual would have several blended emotions with different intensities depending on the respective importance of the desires involved. Actually we believe that in an unfavorable background, joy would yet be triggered (even if it is not expressed due to the presence of more intense negative emotions) and would maybe mitigate these negative emotions. However we do not describe in our model how several emotions triggered at the same time can interact with each other. This could be an interesting continuation of our approach.

4.3 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. They use

[1] This is Lazarus' term for designing his definition of the particular type of relation between the individual and his environment that triggers an emotion (*cf.* Chapter 1, Section 1.3.2).

a local intensity variable called *likelihood*, accounting for the expected probability of the event to occur.

4.3.1 Prospect-based emotions

Definition by OCC (Hope and fear). *An agent feels hope (resp. fear) if he is "pleased (resp. displeased) about the* ***prospect*** *of a desirable (resp. undesirable) event".*

Our formal definition 2 (Hope and fear).

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

Example by OCC. *For example a woman w who applies for a job (p. 111) might feel* ***fear*** *if she expects not to be offered the job ($Expect_w\,\neg$get-job), or feel* ***hope*** *if she expects that she will be offered it ($Expect_w$ get-job).*

An employee e (p. 113) who expects to be fired ($Expect_e\,f$) will feel ***fear*** *if it is undesirable for him ($Des_e\,\neg f$), but not if he already envisaged to quit this job since in this case we can suppose that this is not undesirable for him ($\neg Des_e\,\neg f$).*

4.3.2 Choices of formalization

4.3.2.1 Likelihood

We formalize *likelihood* with the abbreviation $Expect_i$ defined below.

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 the possibility that it could be false". We can notice that if i expects something then he necessarily envisages it:

$$Expect_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi \tag{4.1}$$

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

$$Expect_i\,\varphi \rightarrow \neg Expect_i\,\neg\varphi \tag{4.2}$$

4.3.2.2 Probabilities

The probabilities considered here are debatable. Indeed one can say that on the contrary, one feels hope when something desirable is little probable, like winning Loto. Besides, Ortony *et al.*'s definition of hope is opposite to Lazarus' one in terms of the considered probabilities.

Actually we believe that what determines hope or fear is not only the probability of the expected event, but its expected utility. This expected utility is a function of the probability of a gain, the probability of a loss, and the importance of this gain or loss. For example in playing Loto the expected gain is very huge while the expected loss is small, so the expected utility is positive. Such a situation thus results into a hope emotion whose intensity depends on the exact value of this expected utility. On the contrary when one bets a great amount of money one rather feels fear of losing everything. Indeed, even if the probability to win can be supposed to be the same as in Loto, the expected loss is important so that finally the expected utility is this time negative. In this work we choose to be faithful to the OCC typology, and thus we only consider probability with the same distribution as in the original definition that we formalize.

4.3.2.3 Present or future

One could also object that our definition accounts for the present and not for the future. Actually the object of hope (or fear) is not necessarily about the future. For example an agent can ignore whether his mail has been delivered to the addressee, and thus hope it has been so. However, he has already sent this mail and perhaps it has even already been delivered. One can also hope that there remains some cake when he comes back home.

If the proposition at hand is about future, it suffices to replace φ by $F\varphi$ in the definition. For example to express that i hopes that the rain will stop at some future instant, we can write:

$$Hope_i\, F\neg rain \stackrel{def}{=} Expect_i\, F\neg rain \wedge Des_i\, F\neg rain$$

Moreover we can notice that $Des_i\, F\neg rain$ is a logical consequence of $Des_i\, \neg rain$. Therefore it suffices that i prefers that it does not rain in general and expects that rain will stop at some future instant to deduce that i hopes so.

Thus our definition accounts for any kind of hope, be it directed towards the future, the present or the past.

4.3.3 Discussion

Lazarus does not consider hope and fear as being opposite emotions. He defines hope as "fearing the worst but yielding for better", *viz*. the individual's future expectations are unfavorable but not hopeless. He thus considers that hope is the opposite of despair. He then distinguishes between fright ("concrete and sudden danger of imminent physical harm") and anxiety ("uncertain, existential threat") that involve particular threats (a physical injury for fright, an existential menace to ego-identity for anxiety). Thus according to Lazarus fright is a kind of concrete fear while anxiety is more abstract. We believe that fright and anxiety are particular cases of Ortony *et al*.'s fear emotion, depending on the type of desire that is threatened by the prospected event.

4.4 Confirmation and disconfirmation emotions

These emotions correspond to the confirmation or disconfirmation of a prospect-based emotion (hope or fear). Actually Ortony *et al*. classify them in the same group as hope and fear. But for the sake of readability we expose them in a separate section.

4.4.1 Confirmation emotions

Definition by OCC (Satisfaction and fears confirmed). *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".*

Our formal definition 3 (Satisfaction and fears confirmed).

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

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

Example by OCC. *A candidate who hoped to get a job and finally gets it feels* ***satisfaction****. An employee who feared to be fired feels* ***fear-confirmed*** *when he is.*

4.4.2 Disconfirmation emotions

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

Our formal definition 4 (Relief and disappointment).

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

Example by OCC. *A candidate who hoped to get a job and does not get it feels* ***disappointment****. An employee who feared to be fired will feel* ***relief*** *when he is not fired (*$Bel_e\,\neg f$*).*

4.4.3 Choices of formalization

4.4.3.1 Preservation of expectations

According to our definitions, we can prove (*cf.* Chapter 5) that:

$$Satisfaction_i\,\varphi \leftrightarrow Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi$$

This reformulation makes clearer a problem raised by this definition: hope is never abandoned. Actually if an agent i believes that anytime in the past he has hoped φ, when he finally believes that φ is true he feels satisfied. This is not intuitive since i can have changed his expectations after he hoped φ. For example let's consider someone who prefers sunny days. During the summer they are more probable so he hopes for them, while during the winter they are less probable so he stops hoping for them. Now if the weather is sunny one day during the winter he should rather feel relieved than satisfied, since this is contrary to his current expectations. Nevertheless our definitions also trigger satisfaction. This prevents us from proving that relief and satisfaction about the same event are inconsistent.

A solution would be to use linear temporal logic with $Until$ and $Since$ operators. We could then write for example

$$Relief_i\,\varphi \stackrel{def}{=} Bel_i\,P(\neg Expect_i\,\neg\varphi\ \textit{Since}\ Expect_i\,\varphi) \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$$

and similarly for the other confirmation and disconfirmation emotions. These new definitions allow to entail the intuitive inconsistencies between confirmation and disconfirmation emotions about the same event, and also solve the problem of abandoned hope (or fear). Nevertheless we prefer the linear temporal logic S4.3, because even if it is not expressive enough in this case, it ensures completeness and soundness results. We thus keep our current definitions in spite of this imprecision.

4.4.3.2 Disconfirmation and well-being

From these definitions, we can prove that both confirmation and disconfirmation emotions entail the corresponding well-being emotions. It is intuitive for confirmation emotions but not for disconfirmation emotions, *viz.* the implications $Relief_i\,\varphi \rightarrow Joy_i\,\varphi$ and $Disappointment_i\,\varphi \rightarrow Sadness_i\,\varphi$ are not very intuitive. Indeed disconfirmation emotions rather seem to occur on a return to a normal situation that was expected to change but that finally did not. Actually this normal situation should not necessarily be desirable or undesirable itself, but the logic of Des imposes that it is. So finally this counter-intuitive property is due to the simplification choice that we make to have a normal desire (*cf.* Chapter 3).

4.4.4 Discussion

Lazarus does not define confirmation-based emotions (satisfaction and fears confirmed). They seem to be included in more general emotions like joy and distress. His theory thus seems to be less precise than the OCC typology on this point.

Among disconfirmation emotions Lazarus only defines relief ("a distressing goal incongruent condition has changed for the better or gone away"), but not disappointment. His definition for relief seems to be more generic than Ortony *et al.*'s one. Indeed he does not impose that the individual was feeling fear, but only that he was facing a goal incongruent situation, that in his theory can trigger any negative emotion.

4.5 Fortune-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*, *viz.* i's attitude towards j. We thus have to formalize these variables.

4.5.1 Good-will fortune-of-others emotions

Definition by OCC (Happy for and sorry for). *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.*

Our formal definition 5 (Happy for and sorry for).

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

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

Example by OCC. *Fred (p. 95) feels **happy for** Mary when she wins a thousand dollars, because he believes that this is desirable for her ($Bel_f\ Des_m\ w$) and he has an interest in the well-being of his friends (in particular in this situation: $Bel_f\ Des_m\ w \rightarrow Des_f\ Bel_m\ w$). So we can deduce that it is also desirable for him ($Des_f\ Bel_m\ w$). Moreover it is probable for him that she knows or will know that she won ($Prob_f\ F\,Bel_m\ w$)[2].*

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

4.5.2 Ill-will fortune-of-others emotions

Definition by OCC (Resentment and gloating). *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.*

Our formal definition 6 (Resentment and gloating).

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

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

Example by OCC. *An employee e (p. 99) can feel **resentment** towards a colleague c who received a large pay raise ($Bel_e\,pr$, $Bel_e\,Des_c\,pr$) because he thinks this colleague is incompetent and thus does not deserve this raise ($Des_e\,\neg Bel_c\,pr$).*

*Finally, Nixon's political opponents (p. 104) might have felt **gloating** about his departure from office ($Bel_o\,Bel_{nixon}\,d$, $Bel_o\,Des_{nixon}\,\neg d$) because they thought it was deserved ($Des_o\,Bel_{nixon}\,d$).*

[2] Fred may not feel 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, Fred 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.

4.5.3 Choices of formalization

4.5.3.1 Desirability for other, liking and deservingness

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 global axioms. For example, when John likes Mary this means that if John believes that Mary desires to be rich, then John desires that Mary be rich, or rather: gets to know that she is rich ($Bel_{john}\, Des_{mary}\, rich \rightarrow Des_{john}\, Bel_{mary}\, rich$). Third, we simplify the concept of *deservingness* by identifying "i believes that j deserves φ" and "i desires that j believes φ". Then 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.

4.5.3.2 Should the other agent know about the event?

We also add a weak condition in the definition: it must be at least probable for the agent i feeling the emotion that the other agent j learns about the event at a moment in the future. Ortony *et al.* do not impose in their definition that the other agent should know about this event presumably (un)desirable for him, but we believe that one cannot be happy or sorry for another agent j about something that j does not even know, and thus about what j is not even happy or sad himself.

Other authors (Ochs, Sadek, and Pelachaud, 2007) do not impose this condition in their empathetic emotions, but we believe that this can lead to unsuitable emotions. In this case we are thus more precise than the original definition.

4.5.4 Discussion

Lazarus distinguishes between envy ("wanting what someone else has") and jealousy ("resenting a third party for loss or threat to another's affection" or favor). These two emotions seem to be particular cases of the resentment emotions defined by Ortony and colleagues. Elliott (1992) has implemented the OCC typology and has also refined the resentment emotion into two emotions: envy is resentment over a desired non-exclusive goal while jealousy is resentment over a desired mutually exclusive goal. Once again these are particular cases of the general resentment emotion type defined by Ortony *et al.*, depending on the mutual exclusivity or not of the involved desire.

4.6 Attribution emotions

The emotion types in this group have eliciting conditions focused on the approving of an agent's action. In addition to the appraisal variable of *praiseworthiness* 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 work we only focus on emotions felt by the agent about his own actions, because this variable is to 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[3].

Notation. *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 1. *These emotions are about an action* α *that the agent believes to have influenced the proposition* φ*,* viz. *the agent believes that "if he had not performed action* α*,* φ *would probably be false now". Nevertheless, our language is not expressive enough to represent this counterfactual reasoning, so we make the hypothesis that the agent* i *believes that the parameters* α *and* φ *are linked in this way. The following definitions do make sense only when this is the case.*

4.6.1 Self-agent attribution emotions

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

Our formal definition 7 (Pride and shame).

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

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

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

Example by OCC. *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): $Bel_m\, Done_{m:\alpha}\, \top \wedge Bel_m\, s$[4]. 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*[5] *($Prob_m\, After_{m:\alpha}\, \neg s$).*

*A rich elegant lady l (p. 142) would feel **shame** if caught while stealing clothes in an exclusive boutique ($Shame_l\, (\alpha, \top)$, where α is the action to steal; the result is $\top$ here because this emotion does not depend on the success or failure of the action but on its very performance) because she has performed an action that was unideal for her*[6] *($Idl_l\, \neg Happens_{l:\alpha}\, \top$) and improbable to be performed by her ($Prob_l\, After_{l:\alpha}\, \bot$) due to her social role.*

4.6.2 Other-agent attribution emotions

Definition by OCC (Admiration and reproach). *Emotions involving another agent*[7]*: an agent feels admiration (resp. reproach) towards another agent if he is approving (resp. disapproving) of this agent's praiseworthy (resp. blameworthy) action.*

Our formal definition 8 (Admiration and reproach).

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

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

Example by OCC. *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 to conduce experiments, with the result w of obtaining Nobel-prize-winning*

[4] Actually, she also believes that she influenced this result by her action, *viz.* 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.

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

[6] 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.

[7] When $i = j$, these emotions correspond to the self-agent emotions (*cf.* Theorem 7).

findings) because they internalized this result as ideal[8] *(*$Idl_c\ Happens_{p:\alpha}\ w$*) and difficult (*$Prob_c\ After_{p:\alpha}\ \neg w$*).*

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

4.6.3 Choices of formalization

4.6.3.1 Does the agent appraise the very action or its result?

Our first problem when formalizing these emotions was that there seemed to exist two types of pride, shame, or other attribution emotions: those concerning an action, and those concerning its result. For example one can be proud of his action of participating in the Olympic Games, or can be proud of the result of this action *e.g.* winning the gold medal. For example an agent can also be ashamed of failing in an easy exam (because he failed an easy action) or ashamed of stealing in a shop (because he performed a forbidden action). In a preceding publication (Adam et al., 2006c) we accounted only for emotions concerning the performance of an action and wrote:

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

Then we tried to unify the two types of pride in one single definition. This emotion now concerns both an action α and its result φ. Ortony *et al.*'s notion of *expectation deviation* can account for the two cases: in one case what is expected is that one succeeds in an easy action, in the other case what is expected is that one respects the standards. Both cases can be expressed through our $Expect$ operator.

We can notice that on the contrary we do not capture the attitude of pride (for example someone can be proud of his nationality). Indeed this is a long-lasting attitude that is not triggered by an action, whereas the emotion of pride is triggered by a particular action and then decays.

4.6.3.2 Expectation deviation

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", *viz.* j expects i not

[8] Here, what is ideal is not the very 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.

to achieve φ as a result of his action, for example because it is difficult. The deviation comes from the fact that after the execution of α, j believes that φ is nevertheless true, contrarily to what he expected[9]. 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 to make him feel attribution emotions related to them. This makes it possible for 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 he would if he was forced to wear so.

4.6.3.3 Internalized standards

What is involved in our definitions is not a law (classically represented through a deontic operator $Oblig_s$ indexed by the institution that imposes the law) but an internalized standard. What we call internalized standard is a norm or law that the agent has accepted for himself, that he judges important. Thus our deontic operator Idl_i is indexed by the agent who has internalized this standard and considers it as one of his own ideals. For example a mafia member can kill people without feeling ashamed because he did not internalize the law forbidding to kill.

Indeed we do not feel emotions each time we respect or violate laws or standards, but only when these law are important for us. For example a girl who comes to school with a too short skirt prohibited by the school rules does not feel ashamed about this, because in her mind it is not important to wear appropriate clothes to go to school (actually in her mind this skirt may be appropriate). Each individual thus has his own internalized standards, stemming from the groups to which he belongs, but different from the standards internalized by the other members of these groups. We call these internalized standards the agent's ideals and formalize them with our modal operator Idl_i .

4.6.3.4 Are ideals conscious before the action ?

Finally, we do not impose that the ideal was known at the moment of the action. For example one can feel ashamed about having performed an action when he realize that it was blameworthy, even if he ignored that while performing it. For example, someone who visit Japan can ignore the standards that hold there. Now when he greets someone by shaking hands, he ignores that this is not the standard salutation.

[9] What is unexpected is not just the performance of the action but also its result φ; actually, in the case where φ is $\top$ it is the very performance of the action that is unexpected.

When he later learns that he should have just bowed instead of shaking hands he can feel ashamed about his action. Therefore we do not impose that the standards are conscious at the moment of the action, but only that they are conscious when the agent feels the emotion.

Ideally, we should not impose it either for the probability. For example one can be unaware of the risks when he dives into the water to save a drowning child; he thus feels pride only when he realizes that his action was very brave. But the $Prob_i$ operators are intrinsically epistemic (*viz.* semantically, probable worlds are a subset of possible worlds compatible with the agent's beliefs); so technically it is difficult to avoid consciousness of probabilities in our present setting.

4.6.3.5 Whose is the ideal in other-agent emotions?

In self-agent emotions, the agent who performs the action upholds or violates one of his own ideals. In other-agent emotions, the agent feeling the emotion believes that the other agent's action upholds or violates a standard. The problem is to determine whose is the involved standard.

In this work we consider that an agent i feels an attribution emotion towards another agent j if he believes that j has performed an action that upholds or violates i's own standards (and not j's ones). For example the teacher who reproaches her clothes to a student refers to his own internalized standards: he considers important that students wear appropriate clothes at school. He can feel reproach towards a student even if this student did not internalize this standard *viz.* does not consider it important. That is, the teacher can feel reproach towards the student even if the student does not feel shame about his action. Thus in our definitions the relevant standards are always those of the agent who feels the emotion.

4.6.3.6 About group ideals

The standards stemming from several groups to which an individual belongs can differ. Therefore he may potentially feel different emotions about a same action. For example the school girl may feel ashamed in front of the teacher who reproaches her to wear such clothes, while she would feel pride in front of her friends who admire her trendy skirt. A classical example involves a catholic soldier: the standards in the army impose him to kill enemies while his religious standards forbid him to kill people. When killing an enemy he may thus feel shame in front of a priest but certainly not in front of the other soldiers.

A solution would be to make precise the group that imposes the standard in the operator Idl and thus also in the attribution emotion. An agent would thus feel an attribution emotion towards a given group that imposes the involved standard. For

example:

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

For the sake of simplicity we do not consider such group ideals for now. The study of the properties of this new operator could be subject of later work.

4.6.4 Discussion

Lazarus has a similar definition for pride: "enhancement of one's ego-identity by taking credit for a valued object or achievement, either our own or that of someone or group with whom we identify". He also considers what Ortony *et al*. call *strength of unit* through this notion of identification with a person or a group.

Lazarus distinguishes between guilt ("having transgressed a moral imperative") that he characterizes as self-disgrace, and shame ("a failure to live up to an ego-ideal") that he characterizes as social disgrace. This distinction seems to match the two cases that we envisaged: the shame of failing an action *viz*. of not reaching the expected result matches Lazarus' shame; the shame of performing a prohibited action matches Lazarus' guilt. We have unified these two cases in one single definition.

Besides Lazarus does not define other-agent attribution emotions: admiration and reproach. Once again his theory seems to cover less situations than does Ortony *et al*.'s one, even if he describes some situations more precisely.

4.7 Well-being and attribution composed emotions

4.7.1 Composed emotions

Definition by OCC. *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.*

Our formal definition 9.

$$Gratification_i\,(i{:}\alpha,\varphi) \stackrel{def}{=} Pride_i\,(i{:}\alpha,\varphi) \wedge Joy_i\,\varphi$$

$$Remorse_i\,(i{:}\alpha,\varphi) \stackrel{def}{=} Shame_i\,(i{:}\alpha,\varphi) \wedge Distress_i\,\varphi$$

$$Gratitude_{i,j}(j{:}\alpha,\varphi) \stackrel{def}{=} Admiration_{i,j}(j{:}\alpha,\varphi) \wedge Joy_i\,\varphi$$

$$Anger_{i,j}(j{:}\alpha,\varphi) \stackrel{def}{=} Reproach_{i,j}(j{:}\alpha,\varphi) \wedge Distress_i\,\varphi$$

Example by OCC. *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 (*viz. *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 (making research), 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*[10] *($Des_p\ n$), 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): $Shame_{spy}\ (\omega, d) \wedge Distress_{spy}\ d$.*

4.7.2 Discussion

Lazarus defines anger: "a demeaning offense against me and mine". This is rather in agreement with Ortony *et al.*. However he does not define the other composed emotions of the OCC typology: gratitude, gratification, remorse.

If Ortony *et al.* define remorse, they do not consider regret, and neither does Lazarus. Yet avoiding regrets is a powerful motor of action in decision theory.

4.8 Conclusion

In this chapter we thus formalized twenty emotions of the OCC typology while staying as close as possible to their psychological definitions in (Ortony, Clore, and Collins, 1988). To measure how much close we are to Ortony *et al.*'s definitions we showed for each of our formal definitions that it could capture the author's illustrating example for this emotion. We have also discussed all our necessary interpretations of Ortony *et al.*'s theory and all our choices of formalization on intuitive examples that support them. Finally the next chapter exposes some intuitive properties of emotions that our formal definitions allow to deduce, what validates

[10] This is not always true. For example, one can personally desire not to go to school, while it is ideal to go.

them once more. Nevertheless we have to highlight here some shortcomings in our account that are mainly due to limitations in the expressivity of our modal logic.

4.8.1 Limitations of our account of the OCC typology

4.8.1.1 Intensity of emotions, dynamics and blending

First our modal operators are not graded: they have no associated degree, because it is a complex task to define a semantics for such operators (Laverny and Lang (2004,2005a)). Thus our emotions are not quantitative either, *viz*. they have no intensity. This is an important drawback since the intensity is essential to characterize an emotion (Frijda et al., 1992). This prevents us from formalizing intensity variations in the same type of emotion (for example: irritation, anger, rage), that are yet crucial for an expressive agent. This also prevents us from managing the temporal evolution of emotions, in particular their decay over time. Thus our emotions persist 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 when it is below a threshold the emotion disappears. Finally, we cannot manage emotional blending of several emotions that are simultaneously triggered; Gershenson (1999) proposes an original solution to this issue.

We leave these problems for further work. When we implemented our model (*cf*. Chapter 7) we made some concessions on the semantics in order to have graded emotions. However in this chapter we only give the definition of an emotion but we provide no way of computing its intensity.

4.8.1.2 About intensity variables

As a consequence, we cannot account for what Ortony *et al*. call *intensity variables* that are supposed to influence only the intensity of a given emotion. Actually we can only treat these variables in an all-or-nothing fashion. That is, when such a variable appeared to be crucial, we integrated it in the definition of the emotion, making it indispensable for the very triggering of the emotion. Otherwise we did not take it into account at all in the definition, so that the emotion can exist without it. For example expectation deviation appears in our definitions of attribution emotions while it is only an intensity variable in the OCC typology. A qualitative approximation could be to define several emotions in each of OCC's emotion types, each corresponding to a different degree of intensity and thus involving different intensity variables. But what we want to capture is the general emotion type, and these particular emotions are just variations on it. Thus in this work we only

give one definition corresponding to what we believe to be essential to the defined emotion type.

4.8.1.3 Limitations due to the chosen modal operators

Another problem was to translate the complex concepts that can be expressed in natural language into a limited set of modal operators. It may seem easy to invent operators matching exactly the needed concepts, but it is not. Since we wanted to use our logic to do proofs, we had to ensure correctness and completeness results and thus to make some simplifications. First (*cf.* Chapter 3) we use linear temporal logic $S4.3$ without $Since$ and $Until$ operators whose expressivity could yet have been useful (*cf.* Section 4.4.3.1). Second we have no way of expressing exactly the causal link between an action and its result. The logic $\mathbf{K}_t$ does not provide operators expressing this link. The operator $STIT$ is currently under investigation (Chellas, 1992; Horty and Belnap, 1995b; Troquard, Trypuz, and Vieu, 2006) to fill this gap. Meanwhile we had to approximate this link through making some simplification hypothesis (*cf.* Remark 1). Finally this 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.

Under these simplifications and approximations we managed to formalize twenty emotions of the OCC typology. We then compared on each group of emotions Ortony *et al.*'s typology with Lazarus' theory.

4.8.2 Comparison between Lazarus and OCC

Through this comparison, Lazarus first seems to offer a more precise account of emotions. He uses more complex appraisal variables (for example he distinguishes six types of ego-involvement) to make fine-grained differentiations between some emotions that Ortony *et al.* gather in one single emotion type (*e.g.* guilt and shame, envy and jealousy, fright and anxiety). But meanwhile he neglects some emotions that are considered in the OCC typology (*e.g.* admiration, reproach, remorse, happy for, gloating...) and that seem to be important. Finally we can say that the OCC typology thus seems to be more adapted to design an expressive emotional agent. Indeed what such an agent needs is to express adapted emotions in a great variety of situations, *viz.* it needs to be robust. On the contrary Lazarus privileges precision to robustness, so an agent with this theory could express very fine emotions in some situations, and no emotion at all in some others. His account could thus be useful only in a second step of analysis: if the agent's emotions are not precise enough in some cases we could refine our definitions by looking at Lazarus' theory. We

will thus check if we need such improvements by evaluating our formalization (*cf.* Chapter 7).

To conclude this chapter we now present how we continue this work in the following chapters.

4.8.3 Subsequent work

4.8.3.1 Formal properties of emotions

Provided our proposed formal definitions are accepted, the properties of emotions can now be proved as theorems. In Chapter 5 we expose and prove some properties, particularly relating to causal and temporal links between emotions.

4.8.3.2 Evaluation of the model

These properties along with the formalization of examples from Ortony *et al.*'s book contribute to support the accuracy of our definitions and their faithfulness to the theory. Beyond, we wanted to further assess our model and thus decided to implement it in a BDI agent. This expressive agent was then submitted to the evaluation of human who assessed the relevance and believability of his emotions during a short scenario (*cf.* Chapter 7). This work has given encouraging results and we plan to renew such experiments, in cooperation with psychologists, who could take advantage of such a method to evaluate their theories.

4.8.3.3 Coping

Finally we can notice that this work only concerns the appraisal process, *viz.* the process conducing to the triggering of emotions. But for psychologists, the emotional mechanism not only comprises the appraisal process but also a coping process accounting for the influence of emotions on behaviour. We are currently working on the modeling of this second process in the same BDI formalism. In Chapter 8 we present the current state of this still ongoing research.

To conclude, such a work opens various continuations that can be interesting for several research fields. Embodied conversational agent designers will be interested in giving coping abilities to their agent to make their whole behaviour more emotional and believable (*cf.* Chapter 8). Psychologists may find it interesting to assess psychological theories thanks to an emotional agent (*cf.* Chapter 7). Finally philosophers or logicians may like the possibility to formally prove that some properties hold or not for emotions. This is the object of the next chapter.

Chapter 5

Formal properties of emotions

Deep in the human unconscious is a pervasive need for
a logical universe that makes sense. But the real
universe is always one step beyond logic.
Frank Herbert

5.1 Introduction

As we saw in the state of the art of psychological theories (Chapter 1), emotions are a complex phenomenon that has always been subject to great debates while research got along. Indeed, when concepts are defined informally their definition and properties are always debatable. On the contrary a logical formalization of a concept makes it unambiguous.

In this chapter we show how our sound and complete logical framework (*cf.* Chapter 3) allows to reason about the properties of emotions. This is one step further in disambiguating them, after the formal definitions proposed in Chapter 4 Indeed, provided that our formal definitions are accepted, we can prove some properties of emotions as theorems of our logic that thus are not debatable anymore.

The following sections expose and prove some properties of emotions, particularly referring to the causal and temporal links that exist or do not exist between some of them. Some of these properties go beyond what Ortony *et al*. expose in their book but they remain intuitive. Such a work once more supports the assets of formal reasoning about emotions.

In the proofs, $\mathcal{PL}$ refers to the Propositional Logic, and $\mathcal{ML}$ refers to the principles of normal modal logic (*cf.* Section 3.3.1).

5.2 Prospect-based emotions and their confirmation

5.2.1 New definitions of confirmation and disconfirmation emotions

We wrote our definitions so that they are not redundant. We can prove that as they are defined, confirmation and disconfirmation emotions are equivalent to a past prospect-based emotion and a related belief.

Theorem 1 (New definitions of confirmation and disconfirmation emotions).

$$Satisfaction_i\,\varphi \leftrightarrow Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi \quad \text{(a)}$$

$$FearConfirmed_i\,\varphi \leftrightarrow Bel_i\,PFear_i\,\varphi \wedge Bel_i\,\varphi \quad \text{(b)}$$

$$Relief_i\,\varphi \leftrightarrow Bel_i\,PFear_i\,\neg\varphi \wedge Bel_i\,\varphi \quad \text{(c)}$$

$$Disappointment_i\,\varphi \leftrightarrow Bel_i\,PHope_i\,\neg\varphi \wedge Bel_i\,\varphi \quad \text{(d)}$$

To prove Theorem 1 we need the two following lemmas.

Lemma 1. $Des_i\,\varphi \rightarrow Bel_i\,H\,Des_i\,\varphi$

Proof (of Lemma 1).

1. $Des_i\,\varphi \rightarrow H\,Des_i\,\varphi$ (**from** Pers-Des_i)
2. $Bel_i\,Des_i\,\varphi \rightarrow Bel_i\,H\,Des_i\,\varphi$ (**by** RM-$\Box$ for Bel_i)
3. $Des_i\,\varphi \rightarrow Bel_i\,Des_i\,\varphi$ (**from** 4-MIX2)
4. $Des_i\,\varphi \rightarrow Bel_i\,H\,Des_i\,\varphi$ (**from** 2. and 3.)

Lemma 2. $\vdash Bel_i\,P\,Des_i\,\varphi \rightarrow Des_i\,\varphi$.

Proof (of Lemma 2).

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

Proof (of Theorem 1).
Case of (a).
We will first prove that $Satisfaction_i\,\varphi \rightarrow Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi$.

1. $Satisfaction_i\,\varphi \rightarrow Bel_i\,PExpect_i\,\varphi \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$
(**by** definition of $Satisfaction$)

2. $Des_i\,\varphi \rightarrow Bel_i\,HDes_i\,\varphi$ (**from** Lemma 1)

3. $(PExpect_i\,\varphi \wedge HDes_i\,\varphi) \rightarrow P(Expect_i\,\varphi \wedge Des_i\,\varphi)$
(**from** theorem (3.1) for P)

4. $Bel_i\,(PExpect_i\,\varphi \wedge HDes_i\,\varphi) \rightarrow Bel_i\,PHope_i\,\varphi$
(**by** RM-$\Box$ for Bel_i and definition of $Hope$)

5. $Satisfaction_i\,\varphi \rightarrow Bel_i\,\varphi \wedge Bel_i\,PHope_i\,\varphi$ (**from** 1., 2. and 4.)

We will then prove that $Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi \rightarrow Satisfaction_i\,\varphi$.

1. $Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi \rightarrow Bel_i\,PExpect_i\,\varphi \wedge Bel_i\,PDes_i\,\varphi \wedge Bel_i\,\varphi$
(**from** definition of $Hope$)

2. $Bel_i\,PDes_i\,\varphi \rightarrow Des_i\,\varphi$ (**from** Lemma 2)

3. $Bel_i\,PHope_i\,\varphi \wedge Bel_i\,\varphi \rightarrow Satisfaction_i\,\varphi$
(**from** 1. and 2. by definition of $Satisfaction$)

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

5.2.2 Temporal link from prospect to confirmation

If an agent remembers that at a moment in the past he was feeling a prospect-based emotion about φ, and if he does not ignore anymore if φ is true or false, then our logic entails that he feels the corresponding confirmation emotion.

Theorem 2 (Temporal link from prospect to confirmation).

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

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

Sketch of proof (of Theorem 2). *The proof trivially follows from the new definitions of confirmation and disconfirmation emotions introduced in the previous subsection.*

5.2.3 Inconsistency between confirmation and disconfirmation

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

Theorem 3 (Inconsistency between confirmation and disconfirmation).

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

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

Sketch of proof (of Theorem 3).
The proof immediately comes 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, *viz.* the agent can have expected φ at one moment in the past and $\neg\varphi$ at another moment[1]. We can only prove that these two expectations cannot occur at the same moment (property (4.2)).

5.2.4 Link between confirmation and well-being emotions

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

Theorem 4 (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 (of Theorem 4). *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). □

[1]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 $Relief_i\,\varphi \stackrel{def}{=} Bel_i\,P(\neg Expect_i\,\neg\varphi\ \mathit{Since}\ Expect_i\,\varphi) \wedge Des_i\,\varphi \wedge Bel_i\,\varphi$. Yet, we prefer the linear temporal logic S4.3, because even if it is not expressive enough in this case, it ensures completeness and soundness results.

5.3 Fortunes-of-others emotions

5.3.1 From fortune-of-other emotion to image of other

We can prove that if the agent i feels a fortune-of-other emotion towards another agent j about φ, then he have at least a probability about j feeling the corresponding well-being emotion about φ at a moment in the future.

Theorem 5 (From fortune-of-other emotion to image of other).

$$\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F\, Joy_j\, \varphi \quad \text{(a)}$$

$$\vdash SorryFor_{i,j}\varphi \rightarrow Prob_i\, F\, Distress_j\, \varphi \quad \text{(b)}$$

$$\vdash Resentment_{i,j}\varphi \rightarrow Prob_i\, F\, Joy_j\, \varphi \quad \text{(c)}$$

$$\vdash Gloating_{i,j}\varphi \rightarrow Prob_i\, F\, Distress_j\, \varphi \quad \text{(d)}$$

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

1. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F\, Bel_j\, \varphi \wedge Bel_i\, Des_j\, \varphi$ (**from** definition of $HappyFor$)
2. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, (F\, Bel_j\, \varphi \wedge G\, Des_j\, \varphi)$ (**by** (Pers-Des_i) and (C-MIX))
3. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F(Bel_j\, \varphi \wedge Des_j\, \varphi)$ (**by** property (3.1) for G)
4. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F\, Joy_j\, \varphi$ (**by** definition of Joy)

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

5.3.2 Consequences of fortunes-of-others emotions

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 6 (Consequences of fortunes-of-others emotions).

$$\vdash HappyFor_{i,j}\varphi \wedge \neg Bel_i\, F\, Bel_j\, \varphi \rightarrow Hope_i\, F\, Bel_j\, \varphi \quad \text{(a)}$$

$$\vdash SorryFor_{i,j}\varphi \wedge \neg Bel_i\, F\, Bel_j\, \varphi \rightarrow Fear_i\, F\, Bel_j\, \varphi \quad \text{(b)}$$

$$\vdash Resentment_{i,j}\varphi \wedge \neg Bel_i\, F\, Bel_j\, \varphi \rightarrow Fear_i\, F\, Bel_j\, \varphi \quad \text{(c)}$$

$$\vdash Gloating_{i,j}\varphi \wedge \neg Bel_i\, F\, Bel_j\, \varphi \rightarrow Hope_i\, F\, Bel_j\, \varphi \quad \text{(d)}$$

Proof (of Theorem 6). *Case of (a).*

1. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F Bel_j\, \varphi \wedge Des_i\, Bel_j\, \varphi$
(**from** definition of $HappyFor$)

2. $\vdash HappyFor_{i,j}\varphi \rightarrow Prob_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$
(**by** contraposition of (T-G), and (RM-$\Box$) for Des_i)

3. $\vdash HappyFor_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow$
$Prob_i\, F Bel_j\, \varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$ (**by** $\mathcal{PL}$)

4. $\vdash HappyFor_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow$
$Expect_i\, F Bel_j\, \varphi \wedge Des_i\, F Bel_j\, \varphi$ (**by** definition 1)

5. $\vdash HappyFor_{i,j}\varphi \wedge \neg Bel_i\, F Bel_j\, \varphi \rightarrow Hope_i\, F Bel_j\, \varphi$
(**by** definition of $Hope$)

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

5.4 Attribution emotions

5.4.1 Other-agent emotions towards oneself

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 7 (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 (of Theorem 7).
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). □

5.4.2 Other-agent emotion does not force self-agent emotion

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. For example a fireman who extinguishes a fire may be admired by people watching the scene, while he believes that he is only doing his job and thus does not feel pride.

Theorem 8 (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 (of Theorem 8). *It suffices to find a counter-example,* viz. *a model where the implication is not valid,* viz. *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.*

5.4.3 Link between prospect and attribution emotions

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 (*viz.* about the success or failure of the action w.r.t. the prospected result). We have the same theorem if the agent feeling the emotion is different from the agent performing the action.

For example if an agent is proud of having succeeded in an exam, and if before this exam he envisaged to succeed (and desired to do so), then he was fearing to fail. This example matches the following theorem when α is the action to pass the exam and φ represents the successful result.

Theorem 9 (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 (internalized ideals formalized with our Idl operator) 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.

Proof (of Theorem 9). *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). □

5.4.4 Link between attribution and prospect emotions

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 10 (Link between attribution and prospect emotions). *If α is an action that the agent i believes to influence the proposition φ (*cf. *remark 1), then:*

$$\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \mathit{After}_{i:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Pride}_i\, (i{:}\alpha, \varphi)) \tag{a}$$

$$\vdash \mathit{Hope}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_i\, \neg \mathit{Happens}_{i:\alpha}\, \varphi \rightarrow \mathit{After}_{i:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Shame}_i\, (i{:}\alpha, \varphi)) \tag{b}$$

$$\vdash \mathit{Fear}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_j\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \rightarrow \mathit{After}_{j:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Admiration}_{i,j}(j{:}\alpha, \varphi)) \tag{c}$$

$$\vdash \mathit{Hope}_i\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \wedge \mathit{Bel}_i\, \mathit{Idl}_j\, \neg \mathit{Happens}_{j:\alpha}\, \varphi \rightarrow \mathit{After}_{j:\alpha}\, (\mathit{Bel}_i\, \varphi \rightarrow \mathit{Reproach}_{i,j}(j{:}\alpha, \varphi)) \tag{d}$$

To prove Theorem 10 we need the following lemma.

Lemma 3. $\mathit{Done}_\alpha\, \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \wedge \mathit{Done}_\alpha\, \mathit{Bel}_i\, \varphi \rightarrow \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$

To prove Lemma 3 we need the following lemma.

Lemma 4. *if* $\varphi \rightarrow \mathit{After}_\alpha\, \psi$ *then* $\mathit{Done}_\alpha\, \varphi \rightarrow \psi$

Proof (of Lemma 4).

1. $\varphi \rightarrow \mathit{After}_\alpha\, \psi$ (**by** hypothesis)
2. $\mathit{Done}_\alpha\, \mathit{After}_\alpha\, \varphi \rightarrow \varphi$ (**from** contraposition of (CONV-BH))
3. $\mathit{Done}_\alpha\, \varphi \rightarrow \mathit{Done}_\alpha\, \mathit{After}_\alpha\, \psi$ (**from** 1. by (RK-$\Diamond$) for Done_α)
4. $\mathit{Done}_\alpha\, \varphi \rightarrow \psi$ (**from** 2. and 3.) □

Proof (of Lemma 3).

1. $\mathit{Bel}_i\, \mathit{After}_\alpha\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \rightarrow \mathit{After}_\alpha\, \mathit{Bel}_i\, \varphi$ (**from** (NF-Bel_i))
2. $\mathit{Bel}_i\, \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi \wedge \neg \mathit{Bel}_i\, \mathit{After}_\alpha\, \bot \rightarrow \mathit{After}_\alpha\, \mathit{Bel}_i\, \mathit{Done}_\alpha\, \varphi$ (**by** instantiation of 1.)
3. $\varphi \rightarrow \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi$ (**from** (CONV-AD))
4. $\mathit{Bel}_i\, \varphi \rightarrow \mathit{Bel}_i\, \mathit{After}_\alpha\, \mathit{Done}_\alpha\, \varphi$ (**from** 3. by (RM-$\Box$) 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 4)

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 (of Theorem 10). *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 (RK-$\Diamond$) 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 15, (5-Bel_i) and (C-$\Box$) 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 3)

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). □

5.5 Inconsistencies between some emotions

We can prove several inconsistencies between pairs of emotions.

5.5.1 Polar inconsistencies

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

Theorem 11 (Polar inconsistencies).

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

Sketch of proof (of Theorem 11). *This follows in particular from the rationality axioms (D) for our operators Bel_i, Des_i, $Prob_i$ and Idl_i.*

5.5.2 Non simultaneity of hope and fear

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 12 (Non simultaneity of hope and fear).

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

Sketch of proof (of Theorem 12). *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 (4.2)).*

5.5.3 Inconsistency between good-will and ill-will emotions

Moreover, an agent can not feel simultaneously a good-will and an ill-will emotion towards the same agent about the same issue. This expresses that the other agent is either a friend or an enemy but not both.

Theorem 13 (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 (of Theorem 13). *The proof for cases (a) and (b) follows from the rationality of Des_i. The proof for cases (c) and (d) follows from Lemma 5.* □

Lemma 5. $\neg(Bel_i\, Des_j\, \varphi \wedge Bel_i\, Des_j\, \neg\varphi)$

Proof (of Lemma 5).

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** $\mathcal{PL}$) □

5.5.4 Temporal inconsistency between prospect and confirmation

We can prove that the agent cannot feel simultaneously a prospect-based emotion and a confirmation or disconfirmation emotion about the same object. Actually there must be a temporal step between prospect and confirmation, since the agent gets new beliefs that can be contrary to his expectations.

Theorem 14 (Temporal inconsistency between prospect and confirmation).

$$\vdash \neg(Hope_i\, \varphi \wedge Disappointment_i\, \neg\varphi) \quad \text{(a)}$$
$$\vdash \neg(Hope_i\, \varphi \wedge Satisfaction_i\, \varphi) \quad \text{(b)}$$
$$\vdash \neg(Fear_i\, \varphi \wedge Relief_i\, \neg\varphi) \quad \text{(c)}$$
$$\vdash \neg(Hope_i\, \varphi \wedge FearConfirmed_i\, \varphi) \quad \text{(d)}$$

Sketch of proof (of Theorem 14).
Case of (a). The proof comes from the fact that $Hope_i\,\varphi$ *entails* $Prob_i\,\varphi$ *and thus entails* $\neg Bel_i\,\neg\varphi$*, while* $Disappointment_i\,\varphi$ *entails* $Bel_i\,\neg\varphi$*. The proof is similar for (c).*
Case of (b). The proof comes from the fact that $Hope_i\,\varphi$ *entails* $Expect_i\,\varphi$ *and thus entails* $\neg Bel_i\,\varphi$ *while* $Satisfaction_i\,\varphi$ *entails* $Bel_i\,\varphi$*. The proof is similar for (d).* □

5.6 Other interesting properties

5.6.1 Emotional awareness

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

Theorem 15 (Emotional awareness). *The following formulas*

$$Emotion_i\varphi \leftrightarrow Bel_i\,Emotion_i\varphi$$

$$\neg Emotion_i\varphi \leftrightarrow Bel_i\,\neg Emotion_i\varphi$$

are valid for all $Emotion_i$ *among the twenty emotions defined above.*

Sketch of proof (of Theorem 15). *This follows in particular from the introspection axioms for our operators* Bel_i*,* $Prob_i$*,* $Expect_i$*.*

5.6.2 Emotions and ego-involvement

According to Lazarus (1991), only the situations that are relevant to the individual's well-being can trigger an emotion. If we consider that a situation is relevant to an agent's well-being if it involves one of this agent's desires or values, this is in agreement with the following theorem. Indeed, if the agent has no desire nor ideal, no situation 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 16 (Emotions and ego-involvement). *An agent who has neither desires nor ideals cannot feel any emotion.*

Sketch of proof (of Theorem 16). *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). Composed emotions entail both a desire and an ideal.*

5.7 Conclusion

This work contributes to show the interest of BDI logics to formalize emotions. First, there already exist lots of work about these logics in the agent community and thus such a model is ready to be used in many existing agents. Moreover we showed that provided our formal definitions are accepted, we can prove a lot of intuitive properties of emotions. Only a logical formalism can give such unequivocal results about phenomenons that are not always clearly analysed in the psychological literature.

However, BDI logics are not only a tool for making demonstrations. They also are a powerful mean of describing an agent reasoning. In the next chapter we detail two applications where our model is integrated into a BDI agent, and we show what advantages the agents get from using such a model of emotions.

Part III

Applications and continuations

There is no fire like passion, there is no shark like hatred,
there is no snare like folly, there is no torrent like greed

Buddha

Chapter 6

Applications

I have always wished that my computer
would be as easy to use as my telephone.
My wish has come true.
I no longer know how to use my telephone.
(Bjarne Stroustrup)

6.1 Introduction

Emotional agents get more and more applications (*cf*. Chapter 2) in such various fields as video games, aided-learning, virtual training environments, pedagogical agents, embodied conversational agents, life-like characters... In this chapter, we expose some applications that we explored for our model of emotions. These applications were designed at a time where our model was not as rich as it is now, so they do not cover all the emotions that we defined in Chapter 4. Nevertheless they illustrate that such a logical model can be used not only to reason abstractly about emotions but also to develop emotional agents for concrete applications. Here we expose two of the many possible ones: an intelligent agent who is aware of the user's emotions in order to take care of him in an Ambient Intelligence application; and a conversational agent who expresses his emotions during dialogue to appear more believable in a virtual world for training.

6.2 Application 1: an emotionally aware agent for Ambient Intelligence

Ambient Intelligence is the art of designing intelligent environments, *viz*. environments that can adapt their behavior to their user, to his specific goals, needs... at every moment, in order to insure his well-being in a non-intrusive and nearly invisible way. Here, we want to design a BDI agent that can manage these tasks, and thus has to know about the user's emotions.

Agent designers have investigated different methods to know about the user's emotion when he does not express it directly. Prendinger and Ishizuka (2005) use the results of Picard (1997) to deduce the user's emotion label from monitoring his physiological signals and gaze direction. This method allows to detect in real-time the least changes in the subject's emotions, but it is quite intrusive, disobeying an important principle of ambient intelligence. Another method is explored by Jaques et al. (2004). Their pedagogical agent deduces its pupil's emotion by construing events from his point of view (thanks to a user model), via an appraisal function based on the OCC typology. Carofiglio and Rosis (2005) use the same method in a persuasive agent that takes the user's emotion into account when trying to convince him to adopt a difficult behaviour like changing his eating habits or stopping smoking. This method only provides a subjective view of the user's emotional state, but is quite efficient when associated with a good model of his mental attitudes, and most important it is not intrusive at all. Besides, this method also allows to reason about emotions, for example to understand their causes. It is thus better adapted to the problematic of Ambient Intelligence.

Once the agent knows about the user's emotion he can help him to *cope* with this emotion if it is negative (*cf*. Chapter 8). In psychology (Lazarus and Folkman, 1984) *coping* is the agent's choice of a strategy aiming at suppressing or decreasing a negative emotion that he feels (for example by downplaying or totally suppressing its causes). We consider here that an emotional agent for Ambient Intelligence can help the user in this task.

Finally we believe that for an agent integrated in an Ambient Intelligence System (AmIS), a computational model of emotions is useful in the following cases:

- *(C1)* to compute the user's emotion triggered by an external event (as in (Jaques et al., 2004)) but also:
- *(C2)* to anticipate the emotional effect of its possible actions on the user
  - either to choose an action in order to produce an intended emotional effect *(C2a)*
  - or to decide between actions with comparable physical effects *(C2b)*;

- *(C3)* to understand the causes of an emotion that the user seems to express through his observable behavior. This explanation is made:
  - either directly when all necessary information is known *(C3a)*
  - or through inferring some hypothesis about the user's beliefs *(C3b)*.

  To know the emotion felt by the user and the causes of this emotion is fundamental to act in a really adapted way.

In this first application we use our model of emotions to describe the behaviour of an agent integrated in an AmIS controlling an intelligent house taking care of its dweller by handling his emotions[1]. We demonstrate the power of our framework on five different scenarios corresponding to the five cases (identified above) where the agent needs emotions. In each case we consider the home managing AmIS to be administrated by agent m, who can possibly receive help from other agents of the AmIS. Let h be a human dweller of this house.

6.2.1 Case (C1) : appraisal of an external event from the user's point of view.

By definition, as soon as agent m believes that h's mental state validates the conditions composing a given emotion, m believes that h feels this emotion. For example, let's consider the agent m with the following knowledge base $\mathcal{KB}$ concerning the user h.

Initial Knowledge Base.

- $Bel_m\, Bel_h\, sunny$
- $Bel_m\, Des_h\, sunny$

We can then prove that m believes h to feel joy about the proposition $sunny$.

Deduction. $\vdash \mathcal{KB} \rightarrow Bel_m\, Joy_h\, sunny$

In this case the definition instantly follows from the definition of joy.

Proof.

1. $\vdash \mathcal{KB} \rightarrow (Bel_m\, sunny \wedge Des_m\, sunny)$

[1]Please notice that this application was designed at a time where our model only comprised event-based emotions, so only these emotions are used in the following. Moreover we did not have an account of coping yet, so the possible coping strategies are only mentioned but there are no formal proofs about them.

2. $\vdash \mathcal{KB} \rightarrow Bel_m\ Joy_h\ sunny$ (**from** 1. by definition of joy) □

Thus, if m believes that h believes that the sun is shining (*viz.* $Bel_m\ Bel_h\ sunny$) and m also believes that this is desirable for h ($Bel_m\ Des_h\ sunny$) then by definition m believes that h feels joy about this (*viz.* $Bel_h\ Joy_m\ sunny$).

Once the agent knows about the user's emotion, he can use coping strategies to help the user to cope with it (*cf.* Chapter 8). Here, the considered emotion is positive, so m can aim at maintaining it. Emotions can also be taken into account to modulate the agent's behaviour towards the user, for example the agent should choose the right moment to tell bad news to h.

6.2.2 Case (C2a) : pre-evaluation of the emotional effect of an agent's action on the user, to produce an intended emotional effect.

In some cases, emotional impact can be part of a plan, for example, when the production or removal of some emotion of the addressee of the action accounts for the intended effect (commonly named *Rational Effect* in the agent community (FIPA (Foundation for Intelligent Physical Agents), 2002)). For example, coping strategies can be considered as actions with (indirect) effects on the user's emotions (*cf.* Chapter 8).

Let's suppose that m knows that h feels sadness because it is raining (and thus he cannot take a walk). We can also suppose that agent m desires that h does not feel any negative emotion about any object, in particular he desires that h is not sad about the weather.

Initial Knowledge Base.

- $Bel_m\ Sadness_h\ raining$
- $Des_m\ \neg Sadness_h\ raining$

To reach this intended emotional effect of suppressing this emotion, the agent has several strategies that are close to coping strategies, for example informing h that it is not raining anymore as soon as m learns it[2]. However we do not deal with the planning aspect here so we cannot prove which action m will perform in this situation. The influence of emotions on the agent's actions through the use of *coping* strategies has been subject to ulterior investigations (*cf.* Chapter 8).

[2] m could also try to focus h's attention on something else. This case (yet uncovered here) needs a handling of *activation degrees* accounting for the accessibility of the belief to the conscious. See John Anderson's works in cognitive psychology (Anderson and Lebiere, 1998).

6.2.3 Case (C2b) : pre-evaluation of the emotional effect of an agent's action on the user, to select an action.

It can also be useful to anticipate the emotional impact of an action when various actions with the same relevant informative or physical effect have different emotional effects. These effects are a selection criterion of the action among the other possible actions, if they all allow to reach the physical or informative goal at hand. For example let's suppose that m believes that h desires to play chess and considers three possible opponents: John, Peter and Paul. m also believes that h only expects John to come (and neither Peter nor Paul), while m believes that John cannot come. Finally m believes that h does not believe that he can play chess for now.

Initial Knowledge Base ($\mathcal{KB}_1$ at instant 1).

- *(H1)* $Bel_m\, Bel_h\, G((JohnComes \vee PaulComes \vee PeterComes) \leftrightarrow canPlay)$
- *(H2)* $Bel_m\, Des_h\, canPlay$
- *(H3)* $Bel_h \neg Bel_m\, canPlay$
- *(H4)* $Bel_m\, Bel_h \neg PaulComes$
- *(H5)* $Bel_m\, Bel_h \neg PeterComes$
- *(H6)* $Bel_m\, Expect_h\, JohnComes$
- *(H7)* $Bel_m \neg JohnComes$

m can deduce that h is hoping to play chess.

Deduction (from $\mathcal{KB}_1$).

$\vdash \mathcal{KB} \rightarrow Bel_m\, Hope_h\, canPlay$

Proof (1).
Due to the definition of hope, and since $\mathcal{KB}_1$ trivially entails $Bel_m\, Des_h\, canPlay$ and $Bel_m \neg Bel_h\, canPlay$, it suffices to prove that $\vdash \mathcal{KB} \rightarrow Prob_h\, canPlay$.

1. $\vdash (Prob_h\, JohnComes \wedge Bel_h\, (JohnComes \rightarrow canPlay)) \rightarrow$
   $Prob_h\, (JohnComes \wedge (JohnComes \rightarrow canPlay))$ (**by** C-MIX)
2. $\vdash (Prob_h\, JohnComes \wedge Bel_h\, (JohnComes \rightarrow canPlay)) \rightarrow$
   $Prob_h\, canPlay$ (**by** RM-$Prob_i$)
3. $\vdash Bel_m\, (Prob_h\, JohnComes \wedge Bel_h\, (JohnComes \rightarrow canPlay)) \rightarrow$
   $Bel_m\, Prob_h\, canPlay$ (**by** RM-$\Box$ for Bel_m)

4. $\vdash \mathcal{KB} \rightarrow Bel_m\,(Prob_h\,JohnComes \wedge Bel_h\,(JohnComes \rightarrow canPlay))$

5. $\vdash \mathcal{KB} \rightarrow Bel_m\,Prob_h\,canPlay$ (**from** 4. and 3.) □

m now considers the action α to inform h that John cannot come to play chess with him. We suppose that he believes that this action would not be a complete surprise for h[3]. We also suppose that m is about to inform h that John does not come, and m believes that after this action h will believe this.

Initial Knowledge Base.

- *(H8)* $Bel_m\,Happens_\alpha\,Bel_h\,\neg JohnComes$
- *(H9)* $Bel_m\,\neg Bel_h\,\neg Happens_\alpha\,\top$

Then we can prove that m can deduce that after the performance of this action h will feel disappointed about not being able to play chess anymore.

Deduction.

$\vdash \mathcal{KB} \rightarrow Bel_m\,After_{m:\alpha}\,Disappointment_h\,\neg canPlay$

Proof.

1. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Bel_h\,\neg JohnComes$ (**from** H8 by CD-HA)

2. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Bel_h\,\neg canPlay$ (**from** 1., H1, H4 and H5)

3. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Des_h\,canPlay$
(**from** B2 by Pers-Des_i and GA-MIX)

4. $\vdash \mathcal{KB} \rightarrow Bel_m\,Bel_h\,Expect_h\,canPlay$ (**from** proof (1) and H3)

5. $\vdash \mathcal{KB} \rightarrow Bel_m\,Bel_h\,After_\alpha\,Done_{alpha}\,Expect_h\,canPlay$
(**from** 4. by CONV-AD)

6. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Bel_h\,Done_\alpha\,Expect_h\,canPlay \vee Bel_h\,After_\alpha\,\bot$
(**from** 5. by NF-Bel_i)

7. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Bel_h\,Done_\alpha\,Expect_h\,canPlay$ (**from** 6. by H9)

8. $\vdash \mathcal{KB} \rightarrow Bel_m\,After_\alpha\,Bel_h\,PExpect_h\,canPlay$ (**from** 7. by HB-MIX)

[3] Actually this could be deduced from contextual information. Indeed this is a relevant information so h believes that m should inform him if it is true. Moreover h only expects John to come so he envisages that he could not come.

9. $\vdash \mathcal{KB} \rightarrow Bel_m\, After_\alpha\, Disappointment_h\, \neg canPlay$
(**from** 2., 3. and 8. by definition of $Disappointment$)

Since disappointment is a negative emotion m should avoid to trigger it. Another possible action in this situation is to find an alternative partner so that h can still play chess even if John does not come. This way h would not be disappointed when he learns that John cannot come.

Now m knows that both Peter and Paul might be willing to come to play chess with h too, and must choose the partner that will best fit h's likings. Indeed the action to invite Paul and the action to invite Peter have the same physical effect, *viz.* to allow h to play chess, but they can have different emotional effect. For example let's suppose that m believes that h likes that Paul visits him, but is indifferent to Peter visiting him.

Initial Knowledge Base.

- $Bel_m\, Des_h\, PaulComes$
- $Bel_m\, \neg Des_h\, PeterComes$
- $Bel_m\, Bel_h\, (PaulComes \rightarrow canPlay)$
- $Bel_m\, Bel_h\, (PeterComes \rightarrow canPlay)$

We can then prove in the same way that when m informs h that another chess partner comes home, h revises his beliefs and he feels satisfied about playing chess. Moreover when he learns that the partner is Paul he also feels joy about Paul visiting him, while if he learns that the partner is Peter he is indifferent to that.
If m pursues a strategy of maximizing h's positive emotions then m will rather ask Paul to come than Peter (but this cannot be deduced in our logic since we do not have planning rules).

6.2.4 Case (C3a) observation and explanation of behavior.

Sometimes the observation of the user's behaviour can help the agent to determine what emotion among several possible ones is effectively felt by the user.

For example we suppose that in a sunny morning, m believes that h has to present his work at a meeting today but is not prepared yet. Moreover, his world knowledge tells m that when h is well-prepared he expects his meeting to go well, while when he is not prepared he expects it to go wrong. m also believes that h desires to perform well, and desires the weather to be sunny.

Initial Knowledge Base.

- $Bel_m\ (Bel_h\ prepared \rightarrow Expect_h\ meetingOk)$
- $Bel_m\ (Bel_h\ \neg prepared \rightarrow Expect_h\ \neg meetingOk)$
- $Bel_m\ Des_h\ meetingOk$
- $Bel_m\ Bel_h\ \neg prepared$
- $Bel_m\ Bel_h\ sunny$
- $Bel_m\ Des_h\ sunny$

From this initial knowledge base and his model of emotions, m can deduce two different emotions. You can notice that if m did not know about h's likings he could deduce no emotion from the same information.

Deduction.

- $Bel_m\ Joy_h\ sunny$
- $Bel_m\ Fear_h\ \neg meetingOk$

Now h enters the kitchen and is visibly stressed (we suppose that m has some behavioural laws allowing to interpret h's behaviour and deduce an emotion from it). m can match h's inferred possible emotions with the emotion expressed by his behaviour. Since the stress expressed by h matches the fear of failing the meeting computed by m, m deduces that h is currently more focused on his meeting than on the weather.

Such an information is very useful for m who can now adapt his behaviour to h's needs. For example he may propose to display documents on the kitchen wall to allow h to revise the subject of the meeting while eating breakfast. Moreover such help would also decrease h's fear: once he believes that he is well prepared for the meeting h would not feel fear anymore.

Actually, if m had not known before that h was not prepared, he could yet have inferred some information from the observation of his behaviour. This is detailed in the next case.

6.2.5 Case (C3b) : observation of behavior and explanation hypothesis.

Often the agent's knowledge is not sufficient to determine the object of the user's emotion. In this case, the agent has to observe the user's behaviour and to generate some hypothesis so that the inferred emotion matches the expressed emotion.

For example h comes home in the evening (after his meeting) and m observes that he looks sad. Yet, m does not know why h is sad so we write that h is believed to be sad about an unknown proposition P[4]. To be useful to the user, m must know the object of his sadness so he will now explore his knowledge base and try to find information matching his observation.

We suppose that m knows that h had a meeting, that failing this meeting would be undesirable for him, and that h knows if his meeting has gone wrong or well while m does not.

Initial Knowledge Base.

- *(H1)* $Bel_m\,(Bel_h\,meetingOk \vee Bel_h\,\neg meetingOk)$
- *(H2)* $Bel_m\,Des_h\,meetingOk$
- *(H3)* $\neg Bel_m\,meetingOk \wedge \neg Bel_m\,\neg meetingOk$

No information in this knowledge base allows m to trigger sadness, *viz.* he knows no formula that is undesirable but believed to be true and that would thus match the definition of sadness. Nevertheless he can deduce from it and his emotional definitions the following implications.

Initial Knowledge Base.

- *(H4)* $Bel_m\,(Bel_h\,meetingOk \rightarrow Joy_h\,meetingOk)$
- *(H5)* $Bel_m\,(Bel_h\,\neg meetingOk \rightarrow Sadness_h\,\neg meetingOk)$

One of the two premises is inevitably true due to belief (H1) in the user model. However m cannot know which one is true and thus has to generate a hypothesis. He thus either suppose that the meeting was good or that it was bad[5]. To check the validity of his hypothesis he then uses his observations of h's behaviour. Since h exposes sadness, the validated hypothesis is that he failed his meeting, and the object of the expressed sadness can now be instantiated by an abductive process that we do not specify here.

Note also that abductive inference is not valid, contrarily to deductive inference; therefore there might be other explanations of h's sadness.

Deduction.

- $\vdash \mathcal{KB} \rightarrow Bel_m\,Bel_h\,\neg meetingOk$

[4]This is an expression involving a quantification of the form $Bel_m\,\exists P(Sadness_h\,P)$. As our logic is propositional we suppose that this is handled by some meta-reasoning here.

[5]He could also ask the user whether the meeting was good or not, but this would be more intrusive.

- $\vdash \mathcal{KB} \rightarrow Bel_m\, Sadness_h\, \neg meetingOk$

Thereby m could propose an adapted reaction, for example try to cheer h up or propose him some relaxation services, and particularly avoid to mention the meeting tonight.

6.2.6 Conclusion

This application shows how a BDI model of emotions can allow an agent to reason about the user's emotion. Indeed we have sketched how the five example cases of the introduction can be handled in our framework. Such a reasoning can be useful in Ambient Intelligence as in our example, but also in Human-Computer Interfaces or pedagogical agents. Only a cognitive theory of emotions allows such reasoning about the antecedents of emotions. BDI logics are then particularly adapted to allow an agent to perform this reasoning.

This early application does not handle coping in a formal way yet (in particular we lack $Choice$ and $Intend$ operators) so we have omitted some proofs in our examples. For example in case (C2a) agent m desires that h does not feel sadness about the bad weather. A coping strategy expressed as an action law could allow to infer that as soon as m believes that the rain has stopped, m intends to inform h about that. Such a law can be expressed by the following global axiom: $Bel_m\, Sadness_h\, \varphi \wedge Des_m\, \neg Sadness_h\, \varphi \wedge Bel_m\, \neg\varphi \rightarrow Intend_m\, Bel_h\, \neg\varphi$ reading "if m believes that h feels sad about φ whereas himself knows that φ is wrong, then he will adopt the intention to inform h about this". This intention should lead the agent to inform the user about the weather as soon as it changes, but we do not deal with the planning aspect here. The formalization of coping in our BDI framework has been subject to later and still ongoing work (*cf.* Chapter 8).

6.3 Application 2: emotional dialogue between agents

6.3.1 Introduction

Our logical model of emotions accounts for their cognitive triggering, *viz.* it defines which kinds of mental attitudes constitute the causes of emotions. In this application we focus on how emotions are triggered during dialogue, i.e. we consider emotions triggered by an utterance, both for the speaker and the hearer. However, we do not account for the expression of these emotions (through expressive speech acts for example) nor for their effect on the subsequent dialogue (for example modifying the construction of the dialogue).

A dialogue can be viewed as a sequence of speech acts (Austin, 1962; Searle, 1969; Searle and Vanderveken, 1985) realized by each utterance. Such speech acts

being particular actions they can then be considered as fitting the action branch of the OCC typology. They are thus appraised depending on the conversational norms that the interlocutors must obey (for example Grice's maxims (Grice, 1957)). But this is not sufficient to wholly appraise a speech act, since it conveys some information about the world that must also be appraised. For example, let's imagine a boy who breaks a precious vase and confesses it to his father. If his father only appraises the speech act itself of confessing the fault, he would be proud of his son being brave and honest. But he would certainly also appraise the information "your precious vase is now broken" (what we call the *informational content* here) and feel sad.

Thus receiving a speech act is another way to observe one's environment and to know about relevant stimuli. We consider the reception of a speech act as an indirect perception of a stimulus (described in its informational content, even if the speech act is not an assertive) that can fit any of the three branches of the OCC typology.

Actually we restrict our account here to the events described by assertions and queries. Indeed this application was designed at a time where our model only described eight event-based emotions. Thus we do not account here for the appraisal of a speech act w.r.t. the conversational norms, nor for the appraisal of actions or objects described by a speech act.

In this chapter we will thus analyse an example of emotional dialogue between two avatars[6]. The context is a virtual world for the training of firemen.

6.3.2 Speech act semantics

In the agent communication language (ACL) area, semantics for speech acts is generally in terms of their preconditions and effects. The preconditions describe the conditions that must be true to perform the speech act, and the effects describe the consequences of the speech act on the addressee's mental attitudes. In the ACL area, the most important standard to describe the semantics of speech acts is FIPA-ACL (FIPA (Foundation for Intelligent Physical Agents), 2002). In FIPA-ACL the *Feasibility Preconditions* and *Rational Effect* of various speech acts are described in terms of the agents' mental attitudes and actions. Thus this semantics can be integrated more easily with our formalization of emotions. The *rational effect* describes the desired and rationally-expectable perlocutionary effect of the utterance. Under sincerity and competence hypotheses, the rational effect is true

[6]We have implemented a platform for generating such dialogues between two conversational agents who accompany their speech acts and answer those of the speaker by expressing eight event-based emotions. This platform called GALAAD was presented at a french workshop (Adam and Evrard, 2005)

after each performance of the speech act. For the sake of simplicity, as we are not concerned by the dialogue formalization itself, we make such hypotheses.

We suppose here that the performance of a speech act "activates" the mental attitudes involved in its preconditions and effects. We do not formally handle activation but we suppose that it propagates to mental attitudes concerning the same proposition φ. For example if a fireman informs his chief that there are victims in a fire, this activates the chiefs's related desire that there are no victims. Following the schematic theories of emotions (*cf.* Section 1.2.3.3), we assume that the emotions whose definition involves these activated mental attitudes are then also "activated" by the performance of the speech act. We suppose that the emotions expressed with the speech act are those that are activated like this. Similarly the reception of a speech act adds new mental attitudes or activates old ones in the hearer's knowledge base. These mental attitudes then activate an emotion that corresponds to the agent's reaction to the received speech act.

In the short dialogue that we will formalize in the next section we only need two speech acts of the FIPA library: Inform and QueryIf. The associated effects and preconditions considered in this work are an adaptation of those of FIPA-ACL: actually we weaken the preconditions of these two speech acts to match the needs of our application. The following paragraphs describe this weakened semantics of speech acts.

6.3.2.1 Inform

The speech act used by i to inform j that φ is true is denoted: $\langle i, \mathsf{Inform}, j, \varphi \rangle$.

Precondition. To inform an agent j that φ is true an agent i must believe that φ is true, and must not believe that j knows if φ is true[7]. Finally:

$$Precond(\langle i, \mathsf{Inform}, j, \varphi \rangle) \stackrel{def}{=} Bel_i\, \varphi \wedge \neg Bel_i\, Bel_j\, \varphi \wedge \neg Bel_i\, Bel_j\, \neg\varphi$$

Effect. When an agent i informs an agent j that φ is true the rational effect is that j believes φ. So:

$$\mathit{Effect}(\langle i, \mathsf{Inform}, j, \varphi \rangle) \stackrel{def}{=} Bel_j\, \varphi$$

6.3.2.2 Query-if

The speech act used by i to query j if φ is true is denoted: $\langle i, \mathsf{QueryIf}, j, \varphi \rangle$.

[7]The FIPA semantics also imposes that i must not believe that j is uncertain about φ but we do not impose this condition. Indeed i can confirm what j is uncertain about.

Precondition. To ask j if φ is true, agent i must not know whether φ is true and must believe that j can answer his query[8]. So:

$$Precond(\langle i, \mathsf{QueryIf}, j, \varphi\rangle) \stackrel{def}{=} \neg Bel_i\, \varphi \wedge \neg Bel_i\, \neg\varphi \wedge Bel_i\, (Bel_j\, \varphi \vee Bel_j\, \neg\varphi)$$

Effect. When an agent i asks j if φ is true the rational effect is that j informs i whether φ is true or not. So[9]:

$$\mathit{Effect}(\langle i, \mathsf{QueryIf}, j, \varphi\rangle) \stackrel{def}{=} Done(\langle i, \mathsf{Inform}, j, \varphi\rangle) \vee Done(\langle i, \mathsf{Inform}, j, \neg\varphi\rangle)$$

Now we show how we predict emotions triggered during a simple example of dialogue. We here suppose that the perception of the utterance is sound and complete, and that agents are sincere and competent. Thus, rational effects are systematically produced.

6.3.3 The example dialogue between firemen

We have been involved in a project with ergonomists (*cf.* (El Jed et al., 2005) or (El Jed, 2006)) aiming at developing a virtual training environment for firemen. The following dialogue example is typical of interactions between firemen during an intervention, but it has been simplified for the sake of readability. It involves a fireman f who is asked by his chief c about the situation in a hotel. The chief only knows about the initial situation: he sent a team to this hotel one hour ago because it was blazing. He does not know how things are going now, so he asks his fireman by radio. This scenario is simple but it illustrates the triggering or activation of some emotions after the performance of some speech acts. Here we are only interested in the emotions triggered by the appraisal of the informational content of the speech act so we only have to consider speech acts related to an event.

We use the following propositions: fire means that the hotel is blazing; $\mathit{victims}$ means that there are some victims. We associate each speech act with the emotion expressed by the speaker while performing this speech act (EE), and with the emotion felt by the hearer in reaction to this speech act (TE).

[8]The FIPA semantics also imposes that i must not even be uncertain about φ but we do not impose this condition that seems to be too strong. Indeed i can ask for a kind of confirmation about φ if he is uncertain about it. Moreover the FIPA semantics imposes that i must not believe that j already intends to inform him if φ is true or false. In order to simplify and since we have no modal operator for intention we do not impose this condition. Besides an agent could ask a question even if he believes that the interlocutor already intends to inform him, for example to get a quicker answer.

[9]We note $Done(\alpha) \stackrel{def}{=} Done_\alpha\, \top$ which reads "α has just been done".

- Chief asks: "Did you manage to extinguish the fire?"
  $\alpha_1 = \langle c, \mathsf{QueryIf}, f, \neg \mathit{fire} \rangle$
  $EE : Hope_c \neg \mathit{fire}$
  $TE : Distress_f \mathit{fire}$

- Fireman answers: "No, not yet."
  $\alpha_2 = \langle f, \mathsf{Inform}, c, \mathit{fire} \rangle$
  $EE : Distress_f \mathit{fire}$
  $TE : Disappointment_c \mathit{fire}$

- Chief asks: "Are there any casualties?"
  $\alpha_3 = \langle c, \mathsf{QueryIf}, f, victims \rangle$
  $EE : Fear_c \, victims$
  $TE : Joy_f \neg victims$

- Fireman answers: "No, there are none."
  $\alpha_4 = \langle f, \mathsf{Inform}, c, \neg victims \rangle$
  $EE : Joy_f \neg victims$
  $TE : Relief_c \neg victims$

6.3.4 Analysis of this dialogue

We can suppose that all the firemen prefer that the fire is extinguished and that there are no victims. Moreover the chief fireman considers probable that the hotel is still blazing but that there are no victims since there were no clients at this period. The fireman knows that the hotel is still blazing and that there are no victims.

Initial Knowledge Base ($\mathcal{KB}_c$ of chief c).

- $Des_c \neg \mathit{fire}$
- $Des_c \neg victims$
- $Expect_c \mathit{fire}$
- $Expect_c \neg victims$
- $Bel_c \, (Bel_f \mathit{fire} \vee Bel_f \neg \mathit{fire})$
- $Bel_c \, (Bel_f \, victims \vee Bel_f \neg victims)$

Initial Knowledge Base ($\mathcal{KB}_f$ of fireman f).

- $Des_f \neg \mathit{fire}$

- $Des_f \neg victims$
- Bel_f *fire*
- $Bel_f \neg victims$

We can notice that these two agents already feel emotions in this initial situation: $\mathcal{KB}_c$ entails $Hope_c \neg$*fire* $\wedge$ $Fear_c\, victims$, and $\mathcal{KB}_f$ entails $Distress_f$ *fire* $\wedge$ $Joy_f \neg victims$. But what interests us here is which emotion they will express to accompany the performed speech acts.

6.3.4.1 Step 1: query about the fire

Chief asks: "Did you manage to extinguish the fire?"
$\alpha_1 = \langle c, \mathsf{QueryIf}, f, \neg$*fire*$\rangle$
$EE : Hope_c\, \alpha_1 \neg$*fire*
$TE : Distress_f\, \alpha_1$*fire*

When the chief asks this question he activates the preconditions of its act: $\neg Bel_c$ *fire* and $\neg Bel_c \neg$*fire*. This second belief is part of the definition of $Expect_c$ *fire* that is thus also activated. Then this expectation is part of the definition of the emotion $Fear_c$ *fire* that is thus activated. No other emotion triggered by c's knowledge base is activated by the performance of this speech act.

Activated Mental Attitudes and Emotions (for c after step 1).

- $Expect_c$ *fire*
- $Des_c \neg$*fire*
- $Fear_c$ *fire*

Finally while performing his speech act the chief feels fear that the hotel is still blazing.

The rational effect of this speech act on the fireman is that he informs his chief whether *fire* is true. The precondition for the performance of this speech act is that f believes that φ is true. This mental attitude is thus activated, and consequently the emotion $Distress_f$ *fire* that involves this belief in its definition is also activated.

Activated Mental Attitudes and Emotions (for f after step 1).

- Bel_f *fire*
- $Des_f \neg$*fire*
- $Distress_f$ *fire*

Finally the fireman emotionally reacts to this speech act by expressing his sadness that the fire is not extinguished yet.

6.3.4.2 Step 2: answer about the fire

Fireman answers: "No, not yet."
$\alpha_2 = \langle f, \mathsf{Inform}, c, \mathit{fire} \rangle$
$\triangleright \mathit{Distress}_f\, \alpha_2 \mathit{fire}$
$\triangleright \mathit{Disappointment}_c\, \alpha_2 \mathit{fire}$

As we said in the previous step the performance of the informative speech act about *fire* activates f's sadness that the fire is not yet extinguished.

The rational effect of this speech act on the chief is that he believes that the fire is not extinguished. Since this situation was expected before the fireman's answer, the chief feels fear confirmed. This emotion is not activated by the speech act (the chief did not feel it before), but it is created by the new belief added by the reception of the speech act.

Activated Mental Attitudes and Emotions (of c after step 2).

- $\mathit{Bel}_c\, \mathit{fire}$
- $\mathit{Des}_c\, \neg \mathit{fire}$
- $\mathit{Bel}_c\, \mathit{Done}_{\alpha_2}\, (\mathit{Expect}_c\, \neg \mathit{fire})$

Proof (of the emotion felt by c after step 2).

1. $\mathcal{KB} \vdash \mathit{Bel}_c\, \mathit{Done}_{\alpha_2}\, (\mathit{Expect}_c\, \neg \mathit{fire})$
2. $\mathcal{KB} \vdash \mathit{Bel}_c\, \neg \mathit{Before}_{\alpha_2}\, \neg \mathit{Expect}_c\, \neg \mathit{fire}$ (**from** 1. by definition of Before)
3. $\mathcal{KB} \vdash \mathit{Bel}_c\, \neg H \neg \mathit{Expect}_c\, \neg \mathit{fire}$ (**from** 2. by converse of (HB-MIX))
4. $\mathcal{KB} \vdash \mathit{Bel}_c\, P\, \mathit{Expect}_c\, \neg \mathit{fire}$ (**from** 3. by definition of P)
5. $\mathcal{KB} \vdash \mathit{Bel}_c\, \mathit{fire} \wedge \mathit{Des}_c\, \neg \mathit{fire}$
6. $\mathcal{KB} \vdash \mathit{FearConfirmed}_c\, \mathit{fire}$ (**from** 4. and 5. by def. of $\mathit{FearConfirmed}$)

6.3.4.3 Step 3: query about the victims

Chief asks: "Are there any casualties?"
$\alpha_3 = \langle c, \mathsf{QueryIf}, f, \mathit{victims} \rangle$
$\triangleright \mathit{Fear}_c\, \alpha_3 \mathit{victims}$
$\triangleright \mathit{Joy}_f\, \alpha_3 \neg \mathit{victims}$

As in step 1, this query activates c's mental attitudes related to its preconditions. So similarly we find that this speech act activates c's emotion of hope that there are no victims.

Activated Mental Attitudes and Emotions (of c after step 3).

- $Expect_c \neg victims$
- $Des_c \neg victims$
- $Hope_c \neg victims$

We can notice that this emotion was already felt by c before the performance of his query, but in our sense it is the emotion activated by the performance of this speech act, and thus it is the emotion that c expresses with his query.

When the fireman receives this query, the rational effect is that he answers it. Thus it activates the precondition of his answer *viz.* his belief that there are no victims. This belief matches his desire so finally the fireman's emotional reaction to the received query is joy about the fact that there are no victims.

Activated Mental Attitudes and Emotions (of f after step 3).

- $Bel_f \neg victims$
- $Des_f \neg victims$
- $Joy_f \neg victims$

6.3.4.4 Step 4: answer about the victims

Fireman answers: "No, there are none."
$\alpha_4 = \langle f, \mathsf{Inform}, c, \neg victims \rangle$
$\triangleright Joy_f \, \alpha_4 \neg victims$
$\triangleright Relief_c \, \alpha_4 \neg victims$

As in the previous step, the performance of an informative speech act that there are no victims activates f's joy about this. So f's speech act is accompanied by the expression of $Joy_f \neg victims$.

The rational effect of this speech act on the chief is that he believes that there are no victims. This new belief triggers a new emotion of satisfaction. Indeed this situation matches the chief's past expectations.

Activated Mental Attitudes and Emotions (of c after step 4).

- $Bel_c \neg victims$
- $Des_c \neg victims$
- $Bel_c \, PExpect_c \neg victims$

- $Satisfaction_c \neg victims$

Thus the chief's emotional reaction to the fireman's answer is satisfaction about the received information.

6.3.5 Conclusion

In this application we have shown that our BDI formalization of emotions can be combined with a BDI semantics of speech acts to account for the emotions triggered during dialogue. This account is incomplete for now because this is an early application where we made some simplifications and hypothesis, but it can easily be extended to other cases. Emotions in reaction to the description of an action are triggered similarly to emotions in reaction to the description of an event: the rational effect is in terms of a belief about an action, and this belief can match the definition of an emotion and then activate it. Emotions in reaction to speech acts considered as actions depend on conversational norms that we have not formalized. This could be an interesting future field of research.

Moreover this account is restricted to the expression of emotions during dialogue, but for now these emotions have no influence on the subsequent dialogue. We believe that this influence should be expressed in terms of *coping* strategies, *viz*. the efforts that an individual makes to manage his emotions. A natural continuation of this work would thus be to formalize these coping strategies in our BDI framework ((*cf*. Chapter 8)).

6.4 Conclusion

We have already discussed the advantages of a logical model of emotions. This chapter shows that it is not only useful to do formal proofs of the properties of emotions, but that it is also functioning and ready to be implemented and used in agents for various purposes. We have only explored two of them, but we believe that there exist much more (*cf*. Chapter 2).

Besides these two applications highlight that our model of emotions is incomplete yet. We have formalized the triggering of emotions, but not their influence on behaviour. Thus the previous examples could only be half-formalized. Our first attempt to fill this lack by formalizing *coping* strategies is exposed in Chapter 8.

Chapter 7

Implementation and evaluation

A man should never be ashamed to own he has been
in the wrong, which is but saying, in other words,
that he is wiser today than he was yesterday.
(Alexander Pope, poet)

7.1 Introduction

As we showed it in the introduction of this thesis, the agent community researchers design more and more emotional agents, in particular embodied conversational agents (*cf.* Chapter 2). To ensure the relevance of the expressed emotions in the context of interaction (and thus preserve the believability of the agent) they refer to psychological theories as a basis for their model (*cf.* Chapter 1), and more particularly to the OCC typology (Ortony, Clore, and Collins, 1988)). We believe that the main reason for this wide-spreading is the simplicity of this typology, making it very understandable by computer scientists. However, we observe that from a psychological point of view nothing proves that this theory is better than the other ones. In a context where we try to model agents that are as believable and realistic as possible, this question yet is essential to answer, especially since psychological theories sometimes noticeably disagree about which emotions to consider and how to define them (*cf.* Chapter 1).

In this chapter, we thus want to assess the relevance of the emotions that a BDI agent can express if it is endowed with an emotional model built on our formalization of the OCC typology. To do that we implement our BDI framework of OCC's emotions in a software agent named PLEIAD (*Prolog Emotional Intelligent Agents Designer*)[1]. PLEIAD expresses the emotions that he "feels" in response to

[1]PLEIAD was first presented at the French workshop WACA'2006, *cf.* (Adam, 2006). The results

stimuli sent by the user. From now on, we highlight that PLEIAD experimentally integrates the managing of numerical degrees associated with mental attitudes, and of numerical intensity degrees associated with emotions. We believe that this does not impede this implementation to validate our logical model since the names of the emotions triggered are predicted by the logical model, and only their intensity is computed apart from it. We then run PLEIAD on a short scenario and ask some people to evaluate the relevance of the emotions expressed by the agent w.r.t. the emotions that they would have felt in the same situation, or w.r.t. the emotions that are commonly admissible in this situation. Our aim is:

- to test the predictions of our theoretical model in comparison to the users' expectations;
- to test the predictions of the OCC typology itself in comparison to these expectations.

We will first describe the implementation of agent PLEIAD (Section 7.2), then expose the modalities of our evaluation (Section 7.3), and finally discuss the results of the evaluations and give our conclusions about our model and about its psychological basis (Section 7.4).

7.2 PLEIAD: implementation

In this section we describe the implementation of our logical framework in the PLEIAD agent, in particular the concessions made for the implementation. We also detail the different modules composing our agent's architecture.

7.2.1 Concessions to the logical theory

Compared to the original theory, we made some changes. First, as mentioned in the introduction, we associated degrees with every mental attitudes of the agent, in order to deduce an intensity for each triggered emotion. These numerical values were lacking in our logical framework because it is very hard to give a formal semantics of graduated mental attitudes (*e.g.* (Laverny and Lang, 2005b)). However, they undeniably increase the expressive realism of the agent (*cf.* Section 7.2.5).

Second, we made a concession concerning the completeness of the logic in order to simplify its implementation: we do not handle all logical connectors, so mental attitudes mainly refer to atomic formulas. Moreover we only implemented the more useful axioms so that the implemented axiomatics is not complete. In

of this evaluation will be published in a French journal (Adam, Herzig, and Longin, 2007).

particular, we did not integrated the positive and negative introspection axioms for belief (Axioms 4-Bel_i and 5-Bel_i page 93), probability (Axioms 4-MIX1 and 5-MIX1 page 96) and desire (Axioms 4-MIX2 and 5-MIX2 page 96) to prevent the logical prover to loop infinitely.

Finally, the prover itself is not complete, since it is written in Prolog, so that some valid formulas could not be inferred, but in the standard functioning of our agent this appeared to be sufficient.

7.2.2 Interface

The interface of PLEIAD is shown on figure 7.1. It allows to create an agent by describing his mental attitudes: beliefs, desires, norms, expectations... and to use him in a simulation where he emotionally answers to the stimuli he is sent. The interface is set up of the following items:

- nine frames giving information about the agent and the simulation:
  - the name of the agent;
  - a textual description of his current emotion;
  - a picture expressing his current emotion;
  - the instant of simulation;
  - the agent's beliefs;
  - the agent's ideals;
  - the agent's expectations;
  - the agent's desires;
  - the agent's focus of attention;
- four menus:
  - the Main menu allows to set the debug mode (to see what happens in the Prolog threads called by PLEIAD) and to launch the evaluation scenario;
  - the Designer menu allows to create an agent and to modify him by adding or removing mental attitudes from his knowledge base. This menu will also allow to manage personalities and social roles;
  - the Simulation menu allows to launch or reset the simulation, to send various types of stimuli to the agent, to modify his focus, and to make him use some coping strategies;
  - the View menu allows the user to select additional information to be displayed in satellite windows (all emotions, past beliefs, acquaintances).

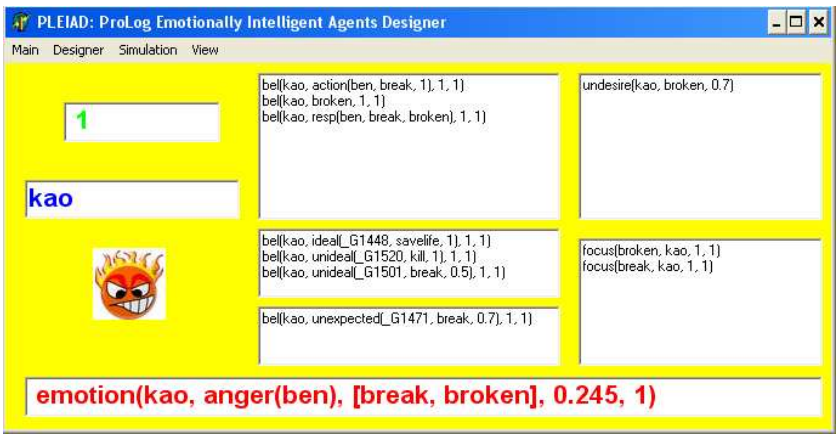

Figure 7.1: PLEIAD interface

7.2.3 Architecture

Our agent architecture uses a knowledge base (KB) containing graduated mental attitudes (*cf.* Section 7.2.5) also associated with an activation degree (or *focus* degree, *cf.* Section 7.2.4). The user can send stimuli (actions or events) to the agent by specifying their name and effects, that are directly added to the agent's beliefs. The perception module is thus transparent for now, since each stimulus is perceived entirely and correctly[2]. A logical prover continuously fills the KB with all the mental attitudes deductible from the mapping of some axioms on its content. For now the axiomatics is incomplete to prevent loops, but this Prolog is sufficient in spite of its incompleteness. An activation managing module generates automatic modifications of focus like temporal decay. The emotional module uses the formal definitions presented in Chapter 4 to deduce all the emotions (associated with an intensity degree) that the agent should "feel" in this situation according to the OCC typology. Finally, a selection module specifies which one of these emotions will be expressed by the expression module[3]. We do not handle facial and bodily animation, so our expression module only displays textual information about the emotion and a smiley to illustrate it.

[2]Later, we could envisage to model the influence of emotions on perception and thus the agent could have an incomplete or distorted perception of stimuli.

[3]The expression module could also take an emotional vector and display a blended facial expression (*e.g.* (Ochs et al., 2005)).

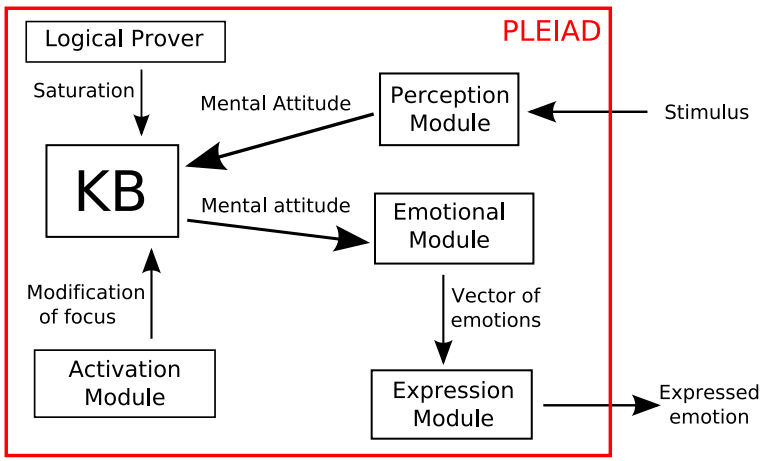

Figure 7.2: PLEIAD architecture

7.2.4 The activation module

According to Anderson (Anderson, 1990; Anderson, 1993; Anderson and Lebiere, 1998; Anderson and Lebiere, 2003; Anderson et al., 2004), activation is a notion determining the accessibility to the conscious of a knowledge unit, or *chunk*. At a given instant, the activation degree of a *chunk* depends on the individual's background (past experiences) proportionately to its subjective utility at this instant. It is computed as the sum of a basic activation and an associative activation. The basic activation represents the past utility of the *chunk viz.* its recency and frequency of use; it decreases logarithmically along time. The associative activation represents the relevance of the *chunk* in the current context (*viz.* w.r.t. the current goal) and depends on the activation degrees of related *chunks*[4]. This theory is implemented in the cognitive architecture ACT-R (Lebiere and Anderson, 1993).

In PLEIAD, we associate a *focus* degree with the mental attitudes (beliefs, desires...) in the agent's KB. The *focus* is a simplified notion of activation in Anderson's sense since it involves no association network linking propositions together, so that its value only depends on what Anderson calls the basic activation. The *focus* degree also represents the availability of a mental attitude for cognitive processes. In our implementation *focus* refer to propositions (noted p), and all mental attitudes referring to the same proposition (*e.g.* $Bel_i\, p$ and $Des_i\, p$) are equally available. We make this reductive choice to simplify the triggering of emotions:

[4]A *chunk* is linked to other *chunks* if it was necessary when these other *chunks* were part of the individual's goal.

once a proposition is available all the terms referring to it in the definition of an emotion are available so this emotion can be triggered.

An empirical initial focus degree is attributed to mental attitudes of the initial knowledge base by the designer of the agent. Then, the perceived stimuli get maximal focus (in PLEIAD it is fixed to 1). Like in ACT-R this focus degree then decreases along time depending on an empirical factor: at each time step the focus degree is multiplied by a given alleviation factor. The focus degree intervenes in the computation of the intensity of an emotion, depending on the activation of its object. Thus this emotional intensity naturally decreases along time while the agent little by little forgets or pay less attention to the object of the emotion.

This way, we can formalize two simple reasoning strategies in Forgas' sense in his work about the *Affect Infusion Model* (Forgas, 1995). According to him, there exists several reasoning strategies among which the individual chooses w.r.t. the context to minimize his efforts. We propose that depending on his current emotion the agent has the choice between two strategies: either he uses his full KB to reason, what is more efficient but also more expensive, or he uses only the most available chunks (those whose focus degree overhauls a given threshold) to possibly get a quicker answer but running the risk to get nothing or something false. However, our model does not explain the choice between these two possible strategies: for now our agent always uses the first strategy.

7.2.5 Emotional intensity

To increase the realism of the emotion expressed by our agent, we wanted to associate this emotion with a gradual intensity degree. Indeed for human beings, to be a little irritated is very different from being really angry. Moreover, the importance of intensity for the expression of emotions and the determination of their influence was already highlighted before, along with the fact that this factor is yet often weirdly ignored.

> "One of the more curious aspects of emotion research indeed is its lack of attention to the fact that emotions vary in intensity. The failure of emotion theorists to address questions concerning emotion intensity is all the more puzzling because intensity is such a salient feature of emotions. Our phenomenal experience acknowledges this fact, as does our behavior and our language; so how is it that our science essentially ignores it? And ignore it, it does."
>
> (Frijda et al., 1992)

According to Lorini and colleagues (Castelfranchi and Lorini, 2003), the degree and dynamics of the emotions arising from the composition of mental attitudes

directly depend on the degree and dynamics of these mental attitudes. They thus associate a subjective certainty degree with the agent's beliefs and an importance degree to his goals, and then compute the intensity of various expectation-based emotions depending on these degrees. In the same way, we also associate degrees with the agent's mental attitudes: a degree of certainty of beliefs, a degree of importance of desires, goals, and standards, a degree of friendship or hostility with acquaintances. We then compute the intensity degree of an emotion as a function (empirically chosen as a product) of the degrees associated with each composing mental attitude and of the activation degree of its object (*cf.* Section 7.2.4). Thus, the intensity of an emotion decreases over time along with the activation of its object. For example the intensity of $Joy_i\,\varphi$ is computed as the product of the degree of belief associated with $Bel_i\,\varphi$, of the degree of desire associated with $Des_i\,\varphi$, and of the activation degree associated with φ. Actually the condition of an emotion can remain true (indeed we made the hypothesis that desires persist) but its intensity decreases along with the agent's focus on its eliciting situation.

We translated the formal definitions of emotions from the BDI logic to Prolog and added the computation of their intensity. All numerical degrees are comprised between 0 and 1. The following example illustrates this process for the joy emotion:

- BDI formula: $Joy_i\,\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Des_i\,\varphi$
- Prolog predicate meaning that the agent I feels joy with degree D about proposition Phi at time T

$$\begin{aligned} cond\_emotion(I, joy, [Phi], D, T) :- & \; infocus(I, Phi, Deg, T), \\ & \; believe(I, Phi, D1, T), \\ & \; desire(I, Phi, D2), \\ & \; D \; is \; D1 * D2 * Deg. \end{aligned}$$

The evaluations confirmed the importance of this intensity degree associated with the emotion, although its numerical expression was not very meaningful for human users.

7.2.6 Emotional expression

The emotional module computes a vector of the emotions induced by the agent's KB and his reasoning principles at a given moment. From the psychological point of view it is fully plausible to feel several emotions simultaneously, even when these emotions are ambivalent *viz.* one is positive while the other is negative (about

this see (Larsen et al., 2004)). PLEIAD thus generates all emotions felt by the agent at a given moment, associated with an intensity degree. These emotions can be visualized in a satellite window through the View menu.

This emotional vector is then intended to be sent to an expression module able to express it, like a facial or body animation engine. Some authors allow their agent to display complex facial expressions resulting from the blending of several emotions (Ochs et al., 2005). They would thus find useful to receive the whole emotional vector generated by PLEIAD. However such an expression module is out of our skills so we choose to give only textual information about the emotion along with an illustrative smiley. In this setting and for the sake of simplicity we make the hypothesis that our agent can only express one emotion at a given instant.

The selection module thus selects the most appropriate emotion in the generated vector. For now this appropriateness only depends on the intensity of the emotion and on its *complexity*. The notion of complexity accounts for the difference in the OCC typology between simple emotions and composed emotions: we consider that the composed emotions are more complex since that convey more information than their constituent emotions. Concretely we choose the object of the most intense emotion and then select the most complex emotion referring to this object. Actually the selection of the most appropriate emotion should depend on many other parameters that we do not account for here: context (some emotions are inhibited in front of some people, for example one should not get angry against his boss), culture, personality...

7.3 Evaluation

Under the hypothesis that our formal model is faithful to a given psychological theory, its faithful implementation allows an evaluation of this underlying psychological theory. Indeed, there exist various psychological theories of emotions, often disagreeing on their definitions. The implementation of our logical model in the agent PLEIAD thus enables to assess not only our BDI model but also its underlying theory: the OCC typology. This section describes the course of evaluations.

7.3.1 Experimental method

PLEIAD provides a test mode where the modifications of the knowledge base are constrained by a scenario predefined by the designer of the agent. The user can only choose between some options at each step of the scenario to influence the continuation *viz*. the agent's KB and thus his/her emotions. At each step an emotion is expressed and the user can fill in a questionnaire about its characteristics. At each

step the user is asked the following questions about this emotion:

- Was it foreseeable? (would the user have felt the same emotion in this situation ?)
- Is it coherent? (can the user understand that someone feels this emotion in the same situation or does he think it to be strange?)

The user can also indicate which emotion he would have felt in this situation if it is different from the generated one, and give some comments.

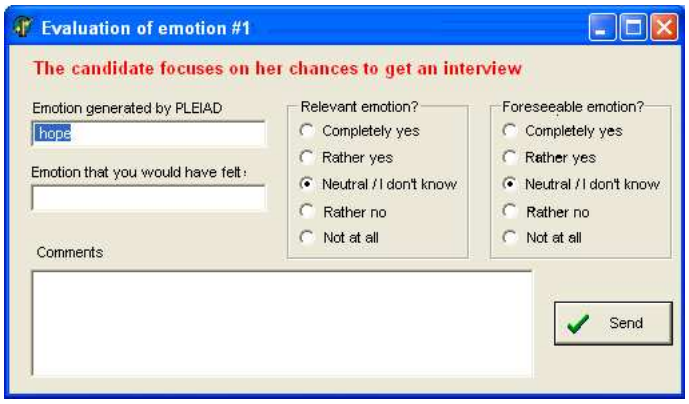

Figure 7.3: PLEIAD questionnaire

We have submitted an interactive scenario to fifteen people and gathered the questionnaires they filled in about the seven emotions triggered during this scenario. Since the judges were not numerous enough to make significative statistics, we mainly exploited their comments to analyse the believability of the expressed emotions and to collect the required improvements.

The scenario that was used during this first evaluation only involves the twelve event-based emotions from the OCC typology. It is detailed in the next paragraph. We intend to make other evaluations with new scenarios involving all the twenty emotions that we formalized.

7.3.2 Scenario

A woman c desires to be hired at an interesting job. The condition to get this job is to get an interview so she also desires to get one. The scenario starts when she sends her curriculum vitæ to the firm, and is set in seven steps.

7.3.2.1 Step 1: chances to get an interview

The first option allows to choose if she considers her CV to be good or bad or if she has no view: the corresponding belief is added to her knowledge base. Moreover she has some world law knowledge allowing to infer her chances to get an interview with such a CV. She also desires to get this interview.

Initial Knowledge Base.

- either $Bel_c\, goodcv$ or $Bel_c \neg goodcv$ or no belief about it
- $Bel_c\, goodcv \rightarrow Expect_c\, getInterview$
- $Bel_c \neg goodcv \rightarrow Expect_c \neg getInterview$
- $Des_c\, getInterview$

A prospect-based emotion is thus generated about her chances to get an interview. If she has no belief about her CV she can infer no expectation and thus no prospect-based emotion.

Deduction. Either $\mathcal{KB} \vdash Hope_c\, getInterview$ or $\mathcal{KB} \vdash Fear_c \neg getInterview$ or $\mathcal{KB}$ infers no emotion.

7.3.2.2 Step 2: convocation for an interview

At the next step, she is convoked for an interview. In our explicit implementation of time, this event increases the current instant value by 1. The desires persist and the agent remembers her past beliefs and expectations.

Initial Knowledge Base.

- either $Bel_c\, PExpect_c\, getInterview$ or $Bel_c\, PExpect_c \neg getInterview$ or no past expectation about the interview
- $Des_c\, getInterview$
- $Bel_c\, interview$

This knowledge base and the definitions of emotions allow to trigger a confirmation-based emotion referring to this convocation.

Deduction. Either $\mathcal{KB} \vdash Satisfaction_c\, getInterview$ or $\mathcal{KB} \vdash Relief_c\, getInterview$ or $\mathcal{KB}$ infers no emotion.

7.3.2.3 Step 3: chances to succeed in the interview

The candidate c then focuses on her chances to succeed in her interview. She desires to succeed. An option allows the user to choose if she is rather optimistic, pessimistic, or neutral.

Initial Knowledge Base.

- either $Expect_c\ succeedInterview$ or $Expect_c\ \neg succeedInterview$ or no expectation
- $Des_c\ succeedInterview$

The third emotion thus refers to the prospect of passing successfully the interview.

Deduction. Either $\mathcal{KB} \vdash Hope_c\ succeedInterview$ or $\mathcal{KB} \vdash Fear_c\ succeedInterview$ or $\mathcal{KB}$ infers no emotion.

7.3.2.4 Step 4: after the interview

At the next step, the user can choose if the woman failed her interview or made it a success. The candidate remembers her past expectations (before the interview) and her desire persists.

Initial Knowledge Base.

- $Bel_c\ \neg succeedInterview$ / $Bel_c\ succeedInterview$
- $Des_c\ succeedInterview$
- $Bel_c\ PExpect_c\ succeedInterview$ / $Bel_c\ PExpect_c\ \neg succeedInterview$

Depending on the configuration of the past expectation and the corresponding current belief, one of the four possible confirmation-based emotions is triggered.

Deduction.

- Either $Satisfaction_c\ succeedInterview$
- or $Relief_c\ succeedInterview$
- or $Disappointment_c\ \neg succeedInterview$
- or $FearConfirmed_c\ \neg succeedInterview$.

7.3.2.5 Step 5: chances to get the job

The applicant then focuses on her chances to get the job. Some world law knowledge allows her to infer expectations about her chances from her beliefs about her success or fail in the interview. Once again she can have no belief and then infer no expectation: in this case she would feel no emotion about getting the job.

Initial Knowledge Base.

- $Bel_c \neg succeedInterview$ / $Bel_c\, succeedInterview$
- $Bel_c\, succeedInterview \rightarrow Expect_c\, cGetsJob$
- $Bel_c \neg succeedInterview \rightarrow Expect_c \neg cGetsJob$

A new prospect-based emotion is thus generated about this new prospected event: getting the job.

Deduction. Either $\mathcal{KB} \vdash Hope_c\, cGetsJob$ or $\mathcal{KB} \vdash Fear_c\, cGetsJob$ or $\mathcal{KB}$ infers no emotion about this.

7.3.2.6 Step 6: results, to be hired or not to be

The user can now choose if the postulant is hired or not.

Initial Knowledge Base.

- $Bel_c\, cGetsJob$ / $Bel_c \neg cGetsJob$
- $Des_c\, cGetsJob$
- $Bel_c\, PExpect_c\, cGetsJob$ / $Bel_c\, PExpect_c \neg cGetsJob$

A confirmation-based emotion among the four possible ones is thus generated about this event.

7.3.2.7 Step 7: about another candidate

Finally, the candidate focuses on another postulant a, who either got the job that she missed or was refused if she got it. c believes that a desires to get the job. The user can then choose if a is a friend, an enemy or a stranger: the corresponding desire[5] is added to c's $\mathcal{KB}$.

[5]We use the same global axioms as in Section 4.5.3.1 to infer the term of the definition corresponding to the desire concerning the other agent.

Initial Knowledge Base.

- $Bel_c \neg cGetsJob \wedge Bel_c\, aGetsJob \wedge Bel_c\, Bel_a\, aGetsJob$ / $Bel_c\, cGetsJob \wedge Bel_c \neg aGetsJob \wedge Bel_c\, Bel_a \neg aGetsJob$
- $Des_c \neg Bel_a \neg aGetsJob$ / $Des_c\, Bel_a \neg aGetsJob$ / no desire about a

A fortune-of-other emotion is then triggered toward the other applicant. If a is a friend it is a good-will emotion, if a is an enemy it is an ill-will emotion, and if a is a stranger (*viz.* a is indifferent to c) c has no desire about him and thus feels no emotion towards him.

Deduction.

- Either $HappyFor_{c,a}\, aGetsJob$
- or $SorryFor_{c,a} \neg aGetsJob$
- or $Resentment_{c,a}\, aGetsJob$
- or $Gloating_{c,a} \neg aGetsJob$
- or $\mathcal{KB}$ infers no emotion.

The next section presents the conclusions that we drew from this evaluation.

7.4 Results

PLEIAD allows us to evaluate our system on three levels.

- First we can evaluate the implementation: is it faithful to our logical model? Does PLEIAD always express the emotions that we expected it to express according to our definitions?
- Second if the implementation is supposed to be faithful, we can evaluate our logical formalization? Are there some cases where the OCC typology says that a given emotion should occur but where our formalization of the situation is such that this emotion is not triggered? Why?
- Finally when our formalization is not responsible for the difference between the emotion expressed by PLEIAD and the emotion expected by the users, we can evaluate the OCC typology itself.

The triggered emotions are globally well accepted by the users, and they consider the agent as being rather believable, even if they found that some emotions were hardly relevant or really aberrant. We thus brought to light several problems, as well in our formalism or our interface as in the OCC theory. We discuss in details these problems in the following paragraphs.

7.4.1 The persistence of emotions seems to be unrealistic

When the received stimulus (the last event that occurred) does not trigger any new emotion, we do not express a neutral answer to this stimulus. Indeed, our selection module chooses an emotion by considering its intensity and complexity, but not its recency. Moreover, emotions persist for some time after their triggering. Thus when no new emotion is triggered by a stimulus, the agent expresses an old emotion that still persists. But this way, he seems to be answering to the new stimulus with an unappropriate emotion, while it is not really the case: actually, he does not answer to this stimulus at all, and just keeps on expressing the same emotion as before.

In a first version of our work (Adam and Evrard, 2005) the agent always expressed the emotion corresponding to the last perceived stimulus, even if it was neutral, but this led to brutal changes in his emotion at every new event. This problem was already highlighted by Frijda et Moffat (Moffat, Frijda, and Phaf, 1993). We thus considered that a possible solution was to express the new emotion only when it was more intense than the current one. But the evaluations showed that it was not very believable. The agent should probably express either a complex blending of several emotions, or an emotional sequence: first the emotion related to the stimulus, and then the previous one if it was more intense.

7.4.2 The status of surprise

Several users said that they would rather feel surprise in some situations with which the OCC typology associates disconfirmation emotions (relief or disappointment) or even well-being emotions. Actually, the emotion of surprise is not described in the OCC typology, while many categorial models of emotions consider it as a basic emotion (*e.g.* (Ekman, 1992b)). Indeed Ortony, Clore, and Collins consider that an emotion is a valenced reaction and thus it should be a valenced feeling. According to them surprise is a "cognitive state" constituent of disconfirmation emotions, that also integrate the realization or not of a desire that creates the valenced reaction. In our work we choosed to faithfully formalize the OCC typology so we did not formalize surprise as an emotion.

If we try to give a BDI definition for surprise it could be as follows:

$$Surprise_i\varphi \stackrel{def}{=} Bel_i\,\varphi \wedge Bel_i\,PExpect_i\,\neg\varphi$$

This reads "agent i believes that φ is now true while he was expecting it to be false". This definition of surprise is a component of our definitions of relief and disappointment, OCC's positive and negative disconfirmation emotions. Thereby relief would be a good surprise while disappointment would be a bad one. However when the unexpected event involves no desire the agent would feel surprise anyway. This is also contrary to Lazarus' view on appraisal: according to him, a stimulus must be relevant to the individual's well-being (*viz.* in our formalization it should match one of the agent's desires) in order to trigger an emotion. Finally the status of surprise as an emotion seems to be at least debatable.

7.4.3 Perception of fortunes-of-others emotions

In some cases the users considered the generated emotion as completely inaccurate. For instance, when the woman learns that she is refused and that the person hired is an enemy, we trigger a resentment emotion, corresponding to the OCC definition: the agent is displeased by an event believed to be desirable for another agent (here the appraised event is the hiring of the enemy). Yet in this situation several users expected the candidate to be angry, while for OCC this emotion is the result of a blending of sadness (here it is the case, the candidate is sad at not being hired) and reproach. According to OCC, reproach arises when the agent disapproves of a blameworthy action, an action that does not respect norms, so it can not arise here, since the enemy has the right to apply for the job, and even the right to get it. According to this definition, the anger felt by the users seems to be wrong.

Lazarus proposes a refinement of the resentment emotion: when others get something that he also wanted and did not get, the individual can feel either envy, or jealousy if this desired resource is mutually exclusive. In the situation where there is just one job to get, the resource is mutually exclusive, and the candidate could feel jealousy towards her rival, friend or enemy. But this is still not anger. Then the difference could be explained by the difficulty to label one's own emotions: were the users able to get the postulant's point of view and determine the emotion that they would have felt? Our next experimentation will be conducted in collaboration with psychologists to answer this problem.

7.4.4 A lack of precision in complex emotions

Another emotion that was not well accepted is "compassion" (*sorry for*) triggered when the candidate gets the job but learns that a friend of hers missed it. By

appraising this event that is undesirable for her friend she feels sorry for him. Yet she is involved in this failure since she is the one who got the job and thus who took it away from him. Some users thus expected her to be embarrassed towards her friend. This emotion is close to the shame of the OCC typology (with a weak intensity) that could arise from the evaluation of her action of applying for the job if this action was blameworthy. There are two reasons explaining that we did not trigger shame in this situation.

First, from the candidate's point of view we choose to represent her hiring as an event, since it was under the responsibility of the recruiter rather than hers, and the recruiter does not belong to the agent's acquaintances. This formalization prevents action-based emotions in the sense of Ortony, Clore, and Collins (like shame or remorse) to occur.

Moreover, we did not consider the action of applying for the job as violating any global standard. Such a standard would make the agent feel shame at being hired in any context, what is not relevant. Instead, we would like the postulant to be ashamed towards her friend for getting the job he also wanted, but to be proud of getting it when she is with her family. One lead is to formalize group-parameterized standards, *viz*. standards that hold for an agent member of a group when he faces this group, but do not hold for the same agent facing another group. For example, a man who is a worshipper but works as a soldier: his work makes him kill people while his religion forbids it, so when he kills an enemy he is ashamed towards his religious community, but not towards other soldiers. Thereby, coming back to our scenario, we could consider that among a group of friends, some implicit standard advises not to try to get something that a friend is desiring too. If this standard held for the candidate, it could explain an emotion of shame in the sense of Ortony, Clore, and Collins towards her friend, while allowing an emotion of pride towards her family.

Finally we can notice that the OCC typology defines remorse as a blending of sadness and shame. In this case the blending is more subtle since the candidate is happy of being hired, but caused a friend to be sad of not being hired, and this action finally makes her sad. We could define a new emotion of embarrassment as a combination of these elements: for j a friend of i,

$$Discomfort_{i,j}(i{:}\alpha, \varphi) \stackrel{def}{=} Shame_i\,(i{:}\alpha, \varphi) \wedge Joy_i\,\varphi \wedge Bel_i\,Sadness_j\,\varphi$$

We could also have a look at Lazarus' theory, that makes a fine-grained distinction between shame and guilt. This could help us to characterize the exact emotion felt by the human users in this complex situation.

7.4.5 About hope, fear, and probability

A chronic problem raised by the users is the confusion between hope and fear. In which case does the agent fear φ and in which case does he hope $\neg\varphi$? The distinction seems to be a matter of probabilities. In some cases people find more realistic to feel hope (resp. fear) when the desirable (resp. undesirable) event is few probable (*e.g.* to hope winning Loto, or to fear being hit while crossing the road; it would seem strange to fear losing or to hope crossing safely). Lazarus defined hope in this way ("fearing the worst but yearning for better" (Lazarus, 1991, p.282)).

Nevertheless Ortony, Clore, and Collins define hope (resp. fear) as the prospect of a desirable (resp. undesirable) event, with the intensity of hope proportional (in particular) to the likelihood of the event. In this setting one would always feel both hope and fear with complementary degrees: for example one would feel a strong hope and a weak fear (resp. a weak hope and a strong fear) when a desirable (resp. undesirable) event is deemed probable. We thus translated the notion of "prospect" with the $Expect$ operator meaning that the agent considers more probable the situation where the event happens than the situation where it does not. This prevents from feeling both hope and fear since their definitions are mutually exclusive. Besides this definition is the opposite (in terms of probabilities) of Lazarus' definition that we used in a previous version of this work (Adam et al., 2006a; Adam et al., 2006b). This disagreement shows the difficulty of distinguishing between these two emotions.

The users also said that these two emotions should not arise when the event is too probable, or not enough probable. This comment is in agreement with Ortony, Clore, and Collins who also notice that if the undesirable event is too probable then the emotion is not fear anymore but rather dread. This restriction is captured by the term $\neg Bel_i \neg\varphi$ in our definition of $Expect_i\, \varphi$, that excludes propositions that the agent believes to be true. Yet it seems that not everybody has the same perception of probabilities: what is important here is the subjective probability of an event that can be different from one person to another. So the problem was that we did not give to the agent the same probabilities of events than the ones "computed" by the users.

7.4.6 About the interface

These evaluations also showed some problems in the ergonomics of the interface. In particular, the object of an emotion is not clearly indicated and users often do not read it and then consider an emotion to be wrong when it is only referring to another object than the one that they are thinking about. This can skew some results, in particular when the emotion is induced by a change of focus that is not

perceived by the user.

Moreover, the intensity of the emotion is given by a numerical degree but this does not seem to be appropriate: this number is not read by the user. We thus changed it to a qualitative degree among three possible semantic values: “few”, “rather”, and “very”. We have simplified the interface (*cf.* Figure 7.4) to give all and only the necessary information in a more comprehensible way, especially for users who do not master Prolog. Thereby, we display “the candidate is very worried about the prospect of not getting an interview” instead of “emotion(employee,fear,not(entretien),0.8,0)”.

Figure 7.4: PLEIAD simplified interface

These simplifications will allow us to conduct new experiments with users who are not computer specialist. We want to conduct these experimentations with the help of psychologists to ensure better expertise. We will also provide more different scenarios. Yet, this first experimentation always gave encouraging results. The expressed emotions are evaluated as globally relevant, and the users’ comments helped us to identify some problems, either in the OCC typology or in our formalization of it.

7.5 Conclusion

In this chapter, we presented the implementation of our logical formalism into a software agent who emotionally answers to stimuli, and we exposed this agent to the critics of human judges to draw conclusions about both the OCC typology and our formalization of it. The results of this first evaluation open plentiful improvement prospects, at least for the part depending on us, *viz.* our own formalization of the OCC typology. Our agent could allow to compare several formalisations of this

typology, provided they are expressed in the same logic. But in addition, it also allows to compare the predictions of several psychological theories of emotions, by formalizing them in the same logic.

In particular, we look toward Lazarus' appraisal theory, which often appeared to be more subtle or more exact during evaluations. However, this theory is far more complex than the OCC typology, partly because it was not designed for an implementation by Artificial Intelligence researchers. It involves complex concepts of responsibility, ego-involvement... that will be hard to formalize in our logic, and that seems to be hard to formalize in a formalism that is not too complex. To formalize Lazarus' theory, Gratch and Marsella (2004a) have created their own complex structure for representing the agent's mental state, by adapting various existing formalisms (*cf.* Chapter 2). A simpler possibility is to represent responsibility through the concept of agency that we can integrate in our BDI logic thanks to the STIT operator (*seeing-to-it-that*, (Horty and Belnap, 1995a)), that we already began to study (*cf.* (Herzig and Troquard, 2006; Broersen, Herzig, and Troquard, 2006)).

However, we have to ask the question of the ratio between the profit taken in terms of expressivity, believability... and the additional costs produced by using such a theory. Is it necessary for an agent to express subtle differences between guilt and shame, or between jealousy and envy, that Lazarus underlines? This highly depends on the application for which the agent is intended (*cf.* Chapter 2). Eventually, the "ideal" emotional theory for these agents could be a compromise between several theories, sometimes using simple but sufficient notions and sometimes using more complex ones for some critical emotions. But in any case, since researchers are currently for different reasons trying to make their agents as believable as possible, we believe that they cannot afford to ignore more complex psychological theories than the classical OCC typology. Moreover, psychology itself may take profit from such researches and in particular from the possibility to evaluate the theories, in order to better understand human emotions.

Chapter 8

Towards a formalization of the coping process

Emotion turning back on itself, and not
leading on to thought or action,
is the element of madness.
(John Sterling)

8.1 Introduction

Emotion takes an increasingly important place in the design of agents for various applications (*cf*. Chapter 2): Ambient Intelligence, believable agents for virtual worlds or video games, emotionally aware agents for tutoring (pedagogical agents), assistance (interfaces agents) or entertainment (virtual companions).

Existing computer science research on emotion (including ours) mainly focuses on the triggering and expression of emotions (Pelachaud et al., 2002; Meyer, 2004). Most of the time, researchers provide an implementation of the OCC typology. Nevertheless this typology only describes *appraisal* (the process assessing the agent's environment to trigger an appropriate emotion), a process that is only part of the human emotional process. Indeed, Lazarus showed that a second process complements *appraisal*: *coping*. In the sense of Lazarus and Folkman (1984), *coping* represents the conscious attempts of the individual to manage threatening stimuli pointed out by intense negative emotions triggered by the *appraisal* process. Very few models of this second process exist (Dastani and Meyer, 2006; Gratch and Marsella, 2004a; Marsella and Gratch, 2003), while several psychological studies endorse the crucial influence of emotions on behaviour (Damasio, 1994; Forgas, 1995).

Yet, as we showed in two applications (*cf.* Chapter 6), agents with *coping* abilities could be useful in several application domains. For example, Embodied Conversational Agents would not only express their emotion with a facial expression or vocal modifications; they would change their whole dialogic behaviour depending on their emotions. Indeed, we believe that some human dialogic behaviours are manifestations of *coping* processes: telling lies, interrupting one's interlocutor, refusing to answer... Standard models of dialogue are unable to explain such kinds of "irrational" but realistic dialogues, since they assume (too) strong and restrictive rationality hypotheses that do not match human reasoning. A model of *coping* strategies could fill this gap. Another example is Ambient Intelligence: in our case study (*cf.* Chapter 6) the intelligent agent who takes care of the user through responding to his emotions actually deploys *coping* strategies for him, to help him manage his negative emotions.

In this chapter we thus propose our first attempt to formalize some coping strategies. This work is still ongoing so the results presented here are preliminary. Our aim is not to provide a full-fledged model of the *coping* process, but rather to disambiguate the concepts involved in the implementation of an emotional agent in order to be able to reason about them. We have already shown the advantages of BDI logics to disambiguate complex concepts (*cf.* Chapter 4) and to reason about their properties (*cf.* Chapter 5). Given the complexity of the emotional process and the limited expressivity of BDI logics compared to natural language, such a model is inevitably simplistic but it offers undeniable assets. We thus build on our BDI framework designed to formalize *appraisal* (*cf.* Chapter 3) and extend it to account for the *coping* process. In particular we need modal operators of choice and intention that are not present in this formalism.

Actually, we adapt the COPE model (Carver, Scheier, and Weintraub, 1989) that proposes a set of fifteen *coping* strategies. We then consider these *coping* strategies as actions whose preconditions and effects are expressed in terms of the agents' mental attitudes (beliefs, desires and intentions). For the sake of simplicity, in this first attempt we restrain to *coping* strategies concerning event-based emotions. We are not interested here in the decision process leading to the choice of one particular *coping* strategy, but only in the effect of the chosen strategy on the agent's mental attitudes. Emotions thus affect the agent's subsequent behaviour in two ways: directly through the choice and application of a *coping* strategy, and indirectly through the modification of the mental attitudes involved in his reasoning.

We start off with an introduction of the psychological concept of *coping* (Section 8.2). We then proceed with the description of the semantics and axiomatic of the new operators that we add in our logical framework (Section 8.3). We will then propose a logical account of some *coping* strategies in this framework (Section 8.4), and illustrate their actual use on an example from a training simulation

for firemen (Section 8.5). Finally we will discuss some existing formalizations of *coping* strategies, namely Gratch and Marsella's EMA agent, Elliott's Affective Reasoner, and Meyer's agent language (Section 8.6).

8.2 The psychological concept of coping

Since the state of the art of this thesis was mainly dedicated to appraisal, we quickly introduce here the psychological concept of *coping* that we want to formalize.

Lazarus and Folkman (1984) quote two origins of the concept of coping: the darwinian theory of stress and control in animals, defining coping as acts controlling aversive situations to lower psychological and physiological perturbations; and psychoanalytic ego psychology, defining coping as realistic and flexible thoughts and acts that solve problems and reduce stress. But for Lazarus, what is important in coping is not the result but the efforts, that differentiate coping from automatic adaptive processes like action tendencies or reflexes. He thus gives a new definition of coping: "constantly changing cognitive and behavioral efforts to manage specific external and/or internal demands that are appraised as taxing or exceeding the resources of the person" (Lazarus and Folkman, 1984). Coping thus has to do with the mastering or minimization of stressful situations. Lazarus then distinguishes two kinds of coping: *problem-focused coping* is oriented toward the management of the problem creating the stress, and is more probable when appraisal indicates a possible solution to this problem; *emotion-focused coping* is oriented toward the regulation of the emotional response to the situation, and is more probable when appraisal indicates no solution to the problem.

Carver, Scheier, and Weintraub (1989) propose the COPE model, that includes a set of fifteen *coping* strategies. In the following we only discuss and formalize the strategies that we judged the most useful for an agent.

- **Active coping** consists in directly acting against the stressor;
- **Seeking emotional social support** consists in trying to get moral support (sympathy, understanding) from other people;
- **Positive reinterpretation and growth** consists in reinterpreting the situation by finding some positive aspects in it;
- **Resignation/acceptance** consists in accepting the reality and move forward;
- **Focus on and venting of emotions** consists in focusing on one's emotion and evacuate it;
- **Denial** is an immature strategy, trying to refuse the reality of the stressor;

- **Mental disengagement** consists in engaging in other activities in order to divert from the stressor.

8.3 Extension of our logical framework

As said before we build on our logical framework (*cf.* Chapter 3) and extend it by a $Choice$ operator (realistic preference) in order to define intention as in (Herzig and Longin, 2004).

8.3.1 Semantics

We add in $\mathscr{R}$ a new structure $\mathscr{C} : AGT \to (W \to 2^W)$ which associates each agent $i \in AGT$ and possible world $w \in W$ with the set $\mathscr{C}_i(w)$ of preferred worlds of agent i in w. All these accessibility relations $\mathscr{C}_i$ are serial, transitive and euclidian.

We associate a modal operator to this mapping: $Choice_i\,\varphi$ reads "agent i prefers worlds where φ is true". The truth condition is standard for this operator: for $i \in AGT$, $w \Vdash Choice_i\,\varphi$ iff $w' \Vdash \varphi$ for every $w' \in \mathscr{C}(w)$.

We have the following additional introspection constraint: if $w \in \mathscr{B}_i(w')$ then $\mathscr{C}_i(w) = \mathscr{C}_i(w')$ ensuring that agents are aware of their choices. We also impose a strong realism constraint: $\mathscr{C}_i(w) \subseteq \mathscr{B}_i(w)$ ensuring that *viz.* all preferred worlds are also compatible with belief.

8.3.2 Axiomatics

The $Choice_i$ operators are defined in the standard KD45 logic (see (Hintikka, 1962; Chellas, 1980; Herzig and Longin, 2004) or Chapter 3). The only relationship between belief and choice is the following axiom of strong realism.

$$Bel_i\,\varphi \to Choice_i\,\varphi \qquad \text{(Real-}Choice_i\text{)}$$

We also have an introspection principle represented by the following mix axiom:

$$Choice_i\,\varphi \leftrightarrow Bel_i\,Choice_i\,\varphi \qquad \text{(Introsp-}Choice_i\text{)}$$

We then follow Herzig and Longin (2004) to define achievement goals and future directed intentions.

Definition 2 (Achievement goal). *Agent i has φ as an achievement goal iff i does not believe that φ is currently true, and in each of his preferred worlds i will believe φ some time. Thus:*

$$AGoal_i\,\varphi \stackrel{def}{=} Choice_i\,FBel_i\,\varphi \wedge \neg Bel_i\,\varphi \qquad \text{(Def}_{AGoal_i}\text{)}$$

Definition 3 (Future directed intention). *Agent i intends that φ iff i has φ as an achievement goal and i does not believe that he will believe φ some day. Thus:*

$$Intend_i\,\varphi \stackrel{def}{=} AGoal_i\,\varphi \wedge \neg Bel_i\,FBel_i\,\varphi \qquad (\text{Def}_{Intend_i})$$

This logic is still sound and complete.

Finally, we consider that we can build on a planning process that we do not detail here. Roughly speaking, if agent i intends that φ be true, and he believes that after α φ will be true, then he intends that α be performed. Thus:

$$\neg Bel_i\,Done_\alpha\,\top \wedge Bel_i\,After_\alpha\,\varphi \wedge Intend_i\,\varphi \rightarrow Intend_i\,Done_\alpha\,\top \quad (\text{PLAN}_{\alpha,\varphi})$$

We can now proceed with formalizing *coping* strategies in this enriched logical framework.

8.4 Formalization of some coping strategies

As we stated in the introduction, we restrain here our account of coping strategies to event-based emotions. Indeed, the definitions of these emotions (*cf.* Chapter 4) all use weak or strong belief and a corresponding individual desire. Our *coping* strategies will modify these mental attitudes to drop the agent's emotion. We believe that action-based emotions may imply other specific *coping* strategies like "shifting responsibility" (Gratch and Marsella, 2004a). Moreover the social ideals underlying these emotions may not be as easy to change as individual preferences. We thus let the study of coping strategies against action-based emotions for future work.

According to Lazarus and Folkman (1984), *coping* strategies only apply to stressful situations, so we are finally only interested in the negative event-based emotions.

Definition 4 (Negative event-based emotion). *An event-based emotion is negative if the involved desire is contradicted or threatened,* viz. *it is in contradiction with a belief or an expectation.*
The negative event-based emotions of the OCC typology are: sadness, fear, fears-confirmed, disappointment, sorry for, and resentment.

However we do not assume a direct correspondence between one emotion and one fixed strategy (contrarily to Dastani and Meyer (2006)). On the contrary we suppose that there is a complex decision process taking the emotion and the context into account to determine the most efficient strategy to drop this emotion. We consider that this process is beyond the scope of our account.

In the next subsections we will formalize some coping strategies as particular actions schemes, and describe their conditions and effects in our logical language.

8.4.1 Formal language

Let $STRA = \{$ActiveCoping, Denial, SeekESupport, Focus&Venting, Resign, PosReinterp, MentalDisengage$\}$ be a subset of action names from ACT corresponding to coping strategies.
Let $EMO^- = \{Distress, Disappointment, Fear, FearConfirmed, SorryFor, Resentment\}$ be the set of negative event-based emotions. Coping strategies can only apply to emotions $E_{i,k}\varphi$ where $E \in EMO^-$. $E_{i,k}\varphi$ is the emotion felt by agent i about φ w.r.t. agent k[1]. Following the psychological literature we call φ the "stressor".
A coping action α is a 4-uple $\langle s, i, E_{i,k}\varphi, \psi\rangle$ where i is the agent applying the coping strategy $s \in STRA$ to the emotion $E_{i,k}\varphi$ thanks to the means ψ. The means is part of the specification of the strategy. When the means is not needed ψ is $\top$: we omit it and we write $\langle s, i, E_{i,k}\varphi\rangle$.

We now give the general action laws of coping strategies.

8.4.2 Action laws

Action laws are made up of executability laws and effect laws. The former describe what must be true before the execution of an action (called the precondition of the action); the latter describe what will be true after the execution of this action (called the effect of the action). In the case of coping actions, preconditions and effects are described in terms of mental attitudes.

To be executable, a coping action must satisfy three conditions: a basic condition (BC) common to all strategies, a control condition (CC) determining which kind of strategy will be applied, and an additional condition (AC) specific to each strategy and constraining its means.

Global axiom 1 (executability laws). *A coping action α executed by agent i is happening next iff all its conditions are satisfied and i prefers that α be performed (Lorini, Herzig, and Castelfranchi, 2006). Thus, for a coping action $i{:}\alpha = \langle s, i, E_{i,k}\varphi, \psi\rangle$*

$$\begin{aligned} Happens_{i:\alpha}\top \leftrightarrow BC(E_{i,k}\varphi) \wedge CC(s, E_{i,k}\varphi) \wedge \\ AC(i{:}\alpha) \wedge Choice_i\, Happens_{i:\alpha}\top \end{aligned} \qquad (\text{EXEC}_\alpha)$$

Definition 5 (basic condition). *Only an agent who feels a negative emotion can cope with it. Thus the basic condition to use any strategy is that the agent believes to feel a negative emotion $E_{i,k}\varphi$ where $E \in EMO^-$:*

$$BC(E_{i,k}\varphi) \stackrel{def}{=} Bel_i\, E_{i,k}\varphi \qquad (\text{Def}_{BC})$$

[1] k is not necessarily different from i. When i is k we will sometimes write $E_i\,\varphi$.

According to Lazarus and Folkman (1984) there are two types of strategies: problem-focused ones, that he considers more likely against a controllable stressor, and emotion-focused ones, more likely otherwise. We capture this distinction in an all-or-nothing way: if the stressor is controllable the agent will only use problem-focused strategies, whereas if it is not he will only use emotion-focused ones.

We consider that a stressor φ is controllable by agent i iff i envisages a possibility to change in the future the fact that φ is true. Conversely, the stressor is uncontrollable iff agent i believes that henceforth he will believe φ to be true. This distinction is not exhaustive: the problem can be neither fully controllable nor fully uncontrollable. The following control condition specifies how the agent i selects a type of coping strategy (problem-focused or emotion-focused).

Definition 6 (control condition). *Thus the control condition of problem-focused coping strategies is that the object of the emotion is controllable; and the control condition of emotion-focused strategies is that the object of the emotion is uncontrollable*[2]*. So:*

$$CC(s, E_{i,k}\varphi) \stackrel{def}{=} \begin{cases} \neg Bel_i \neg F Bel_i \neg\varphi & \textit{if} \quad s = \mathsf{ActiveCoping} \\ Bel_i\, G Bel_i\, \varphi & \textit{else} \end{cases} \qquad (\text{Def}_{CC})$$

We can prove that both control conditions are mutually inconsistent. Thus this condition is really a choice criterion that allows the agent to determine which category of coping strategies he can use.

To select a particular coping action in the selected category, the agent still has to check some additional conditions that are specific to each coping strategy. These conditions are specified in the next section.

Notation. *We note $AC(\alpha)$ the additional condition of the coping action α and $Effect(\alpha)$ its effect.*

Global axiom 2 (effect laws). *Effect laws are defined by instances of the following effect laws scheme:*

$$After_{\alpha}\, Effect(\alpha) \qquad (\text{EFFECT}_{\alpha})$$

where $Effect(\alpha)$ denotes the effect of the action α.

In the rest of this section we describe the additional condition and the effect of each coping action that we define. Thus action laws will be completely defined for all coping actions.

[2]Since the definitions of prospect-based emotions do not involve beliefs but only expectations, this strategy is not applicable to fear.

8.4.3 Formalization of coping actions

We have here selected seven strategies among those of the COPE model (Carver, Scheier, and Weintraub, 1989). We recall that a coping action consists in a coping strategy associated with an optional constraint type of means, instantiated by the decision process depending on the specific context. It is important to understand that a coping strategy with a different type of means makes no sense in our formalism.

Active coping

For the agent to apply an ActiveCoping strategy, he must believe that there is at least one possibility to change (in the future) the fact that φ is true (this matches the control condition). Since the agent's goal when applying such a strategy is to make φ false, it is not necessary for him to apply it if he considers that φ could become false later without acting for that. So the agent will use active coping only if he believes that there exists at least one possibility for him that henceforth he never believes φ to be false. This sets up the additional condition of this strategy. The effect of its application is that the agent adopts the realistic preference that in the future he will believe φ to be false. Thus:

$$AC(\langle \mathsf{ActiveCoping}, i, E_{i,k}\varphi \rangle) \stackrel{def}{=} \neg Bel_i \neg G \neg Bel_i \neg \varphi$$
$$\mathit{Effect}(\langle \mathsf{ActiveCoping}, i, E_{i,k}\varphi \rangle) \stackrel{def}{=} Choice_i\, F\, Bel_i \neg \varphi$$

We can prove that $After_{\langle \mathsf{ActiveCoping}, i, E_{i,k}\varphi \rangle}\, (Bel_i\, \varphi \rightarrow Intend_i\, \neg\varphi)$. (That is: if i still believes φ to be true, he will act in order to make it false.)

Denial

Denial operates by refusing the reality of the stressor, supporting this assumption by some proof; thus the means of denial is a formula ψ which is believed to entail $\neg\varphi$ (*viz.* $Bel_i\, G(\psi \rightarrow \neg\varphi)$). These constraints on the formula ψ constitute the additional condition of denial. The effect of this strategy is to add the belief that ψ is true in the purpose of deducing $\neg\varphi$. Thus, formally:

$$AC(\langle \mathsf{Denial}, i, E_{i,k}\varphi, \psi \rangle) \stackrel{def}{=} Bel_i\, G(\psi \rightarrow \neg\varphi)$$
$$\mathit{Effect}(\langle \mathsf{Denial}, i, E_{i,k}\varphi, \psi \rangle) \stackrel{def}{=} Bel_i\, \psi$$

We can easily prove that: $After_{\langle \mathsf{Denial}, i, E_{i,k}\varphi, \psi \rangle}\, Bel_i \neg\varphi$, so after the performance of the denial strategy agent i is not stressed anymore.

Seeking emotional support

SeekESupport (*viz.* seeking emotional support) operates by looking for the compassion of a friendly agent j who can see the situation; thus its means is $Bel_i\ Bel_j\ \varphi$, and its condition is that the agent j is believed to be friendly. By friendly we mean that agent j should dislike that his friend i believes something undesirable for him (*viz.* $Bel_i\ Des_j\ \neg Bel_i\ \varphi$): this is the additional condition of this strategy. Its effect is that agent i, believing that j is friendly, will adopt the intention to obtain his compassion[3] (*viz.* $Intend_i\ SorryFor_{j,i}\varphi$). To achieve this intention i may have to communicate and explain his emotion to j.

$$AC(\langle \mathsf{SeekESupport}, i, Sadness_i\varphi, Bel_i\ Bel_j\ \varphi\rangle) \stackrel{def}{=} Bel_i\ Des_j\ \neg Bel_i\ \varphi$$

$$\mathit{Effect}(\langle \mathsf{SeekESupport}, i, Sadness_i\varphi, Bel_i\ Bel_j\ \varphi\rangle) \stackrel{def}{=} Intend_i\ SorryFor_{j,i}\varphi$$

Focus on and venting

Focus&Venting operates by looking for the attention of any agent j attending the situation; thus its means is the same as for the previous strategy (*viz.* $Bel_i\ Bel_j\ \varphi$). This strategy is similar to seeking emotional support, but it has no conditions and can be applied to any emotion. Its effect is that the agent i adopts the intention to communicate his emotion to j. This effect is a little weaker than the previous one since i is not sure to obtain compassion from an agent who is not believed to be friendly.

$$AC(\langle \mathsf{Focus\&Venting}, i, E_{i,k}\varphi, Bel_i\ Bel_j\ \varphi\rangle) \stackrel{def}{=} \top$$

$$\mathit{Effect}(\langle \mathsf{Focus\&Venting}, i, E_{i,k}\varphi, Bel_i\ Bel_j\ \varphi\rangle) \stackrel{def}{=} Intend_i\ Bel_j\ E_{i,k}\varphi$$

Resignation

Resign (*viz.* resignation) has no particular means, nor any condition. It is the simplest strategy, consisting in accepting the situation. Actually, to simulate how the agent can get used to the situation over time, we make him drop his contradicted desire[4] immediately. This is an approximation of a long-term process, which leads to the disappearance of the negative emotion.

$$AC(\langle \mathsf{Resign}, i, E_{i,k}\varphi\rangle) \stackrel{def}{=} \top$$

$$\mathit{Effect}(\langle \mathsf{Resign}, i, E_{i,k}\varphi\rangle) \stackrel{def}{=} \neg Des_i\ \neg\varphi$$

[3] $SorryFor_{i,j}\varphi \stackrel{def}{=} Bel_i\ \varphi \wedge Prob_i\ F Bel_j\ \varphi \wedge Bel_i\ Des_j\ \neg\varphi \wedge Des_i\ \neg Bel_j\ \varphi$ (*cf.* Chapter 4)

[4] In all negative event-based emotions $E_{i,k}\varphi$ except fortunes-of-others ones, the contradicted desire is $Des_i\ \neg\varphi$. In fortunes-of-others emotions the contradicted desire is $Des_i\ \neg Bel_j\ \varphi$.

We can prove that after execution of this strategy, agent i no longer believes that he feels the negative emotion $E_{i,k}\varphi$. Formally, $After_{\langle \mathsf{Resign}, i, E_{i,k}\varphi \rangle} \neg Bel_i\, E_{i,k}\varphi$ is provable. Note that such an abandoning of a desire requires to drop our hypothesis that desires are eternal (Axiom (Pers-Des_i) page 97).

Positive reinterpretation

PosReinterp (*viz.* positive reinterpretation) operates by finding a positive aspect in the stressor; so its means is a formula ψ such that ψ is a not undesirable consequence of the stressor. This constraint on the formula ψ sets up the additional condition of positive reinterpretation (*viz.* $\neg Des_i\, \neg\psi$). The effect of this strategy is that the formula ψ becomes desirable. Thus the negative emotion about φ will be replaced by a positive one about ψ.

$$AC(\langle \mathsf{PosReinterp}, i, E_{i,k}\varphi, \psi \rangle) \stackrel{def}{=} Bel_i\, G(\varphi \rightarrow \psi) \wedge \neg Des_i\, \neg\psi$$

$$\mathit{Effect}(\langle \mathsf{PosReinterp}, i, E_{i,k}\varphi, \psi \rangle) \stackrel{def}{=} Des_i\, \psi$$

We can prove that $Bel_i\, \varphi \rightarrow After_{\langle \mathsf{PosReinterp}, i, E_{i,k}\varphi, \psi \rangle}\, Joy_i\, \psi$. In other words, if agent i believes φ then after the execution of this strategy he believes ψ and feels joy about it. We notice that we need the full belief about φ to make this strategy efficient, so it will not work on prospect-based emotions (*viz.* fear). Again, note that such a coping strategy conflicts with our simplifying hypothesis that desires are persistent.

Mental disengagement

MentalDisengage (*viz.* mental disengagement) operates by engaging in an action to take mind off stressor; thus its means is $Happens_{i:\alpha}\, \varphi$, and its condition is that the effect ψ of this action α is believed to be false for now but is desirable for i. The effect of this strategy is that the agent adopts the intention to perform this disengaging action. As a consequence this will trigger in the future a positive emotion about the effect of this action.

$$AC(\langle \mathsf{MentalDisengage}, i, E_{i,k}\varphi, Happens_{i:\alpha}\, \top \rangle) \stackrel{def}{=}$$
$$Bel_i\, \neg\psi \wedge Bel_i\, After_\alpha\, \psi \wedge Des_i\, \psi$$

$$\mathit{Effect}(\langle \mathsf{MentalDisengage}, i, E_{i,k}\varphi, Happens_{i:\alpha}\, \top \rangle) \stackrel{def}{=} Intend_i\, Happens_{i:\alpha}\, \top$$

We can prove that after this strategy the agent feels joy about ψ, *viz.* formally $After_{\langle \mathsf{MentalDisengage}, i, E_{i,k}\varphi, Happens_{i:\alpha}\, \top \rangle}\, Joy_i\, \psi$.

In the next section we illustrate our formalism on an example showing for each strategy how its means is instantiated, how its conditions are verified, and how it influences the agent's mental attitudes and emotions.

8.5 Application on an example

We consider the agent m who is the manager of a hotel. We use our definitions (*cf.* Chapter 4) to compute his emotions in a given situation, and then show how the strategies formalized in the previous section work against his negative emotions.

8.5.1 Initial situation 1

For the agent to use a problem-focused coping strategy, we need to place him in a situation where the stressor is believed to be controllable, *viz.* not definitive. We thus consider the manager of the hotel discovering that a fire has started in his hotel ($Bel_m\ burning$), and we suppose that:

- his hotel could be destroyed

$$Prob_m\ destroyed \tag{H1}$$

- it is undesirable for him

$$Des_m\ \neg destroyed \tag{H2}$$

- he envisages a possibility that his hotel will not be destroyed by the fire

$$\neg Bel_m\ \neg F\,Bel_i\ \neg destroyed \tag{H3}$$

According to our definitions (H1) and (H2) entail that the manager feels fear about the destruction of his hotel and he is conscious of this emotion.

$$Fear_m\ destroyed$$
$$Bel_m\ Fear_m\ destroyed$$

By definition of EMO^-, $Fear$ is a negative emotion. So according to (Def_{BC}), this knowledge base entails that $BC(Fear_m\ destroyed)$ holds: the agent can use any coping strategy against this emotion. Moreover (H3) entails that $CC(\mathsf{ActiveCoping}, Fear_m\ destroyed)$ hold so m may use only problem-focused coping strategies. Among these strategies we only formalized active coping.

8.5.2 Initial situation 2

For the agent to use emotion focused strategies, we now have to send him a stressor that he cannot control. We thus consider that:

- his hotel is destroyed by the fire and this is undesirable for him

$$Bel_m\, GBel_m\, destroyed \tag{H4}$$

$$Des_m\, \neg destroyed \tag{H5}$$

(H5) entails [5] that $Des_m\, \neg G\, destroyed$. Following our definitions of emotions the manager feels sadness about this destruction and he is aware of that:

$$Sadness_m\, destroyed$$
$$Bel_m\, Sadness_m\, destroyed$$

Since $Sadness$ is a negative emotion, it follows from this that the basic condition $BC(Sadness_m\, destroyed)$ holds. Moreover (H4) entails that the control condition Def_{CC} is true for any emotion-focused strategy $s \in STRA \backslash \{\mathsf{ActiveCoping}\}$. The manager may thus apply emotion-focused coping strategies. We will now discuss the context in which each strategy can be applied, and its effects on the manager's sadness.

8.5.2.1 Denial

This strategy consists in finding an argument to support that the stressor φ (believed to be true for now) is actually false. In this situation, the manager wants to deny the destruction of his hotel and thus searches an argument. The manager believes that if the firemen extinguished the fire in time, his hotel is not destroyed (*viz.* $Bel_m\, G(extingInTime \rightarrow \neg destroyed)$. The additional condition for the manager to deny the destruction of his hotel through the hypothesis $extingInTime$ thus holds:

$$AC(\langle \mathsf{Denial}, m, Sadness_m destroyed, extingInTime \rangle) \stackrel{def}{=} Bel_m\, G(extingInTime \rightarrow \neg destroyed)$$

Thus the manager may apply this particular strategy in this situation. If he executes this coping action, the effect is to reinforce his weak belief that the fire was extinguished in time:

$$Effect(\langle \mathsf{Denial}, m, Sadness_m destroyed, extingInTime \rangle) \stackrel{def}{=} Bel_m\, extingInTime$$

[5] From Axiom (T-G) (page 94) and modal principles for Des_m

Thus he can deduce that his hotel is not destroyed ($Bel_m \neg destroyed$). He is thus denying the reality, trying to convince himself that the firemen must have extinguished the fire in time and saved his hotel. As an immediate effect, his sadness will disappear since he no longer believes that the stressor is real. Then he will probably have to use some strategies to prevent himself from seeing the reality, like avoiding talking to people and looking at his hotel for a while.

8.5.2.2 Seeking emotional support

In the same initial situation, we suppose that one of the manager's friends j passes by the hotel. As he is a friend of m, he is believed to dislike that something undesirable occurs to the manager, in particular the destruction of his hotel (*viz.* $Bel_m\, Des_j \neg Bel_m\, destroyed$). The additional condition for m to seek emotional support from agent j seeing the scene (the means of the strategy is $Bel_m\, Bel_j\, destroyed$) thus holds.

$$AC(\langle \mathsf{SeekESupport}, m, Sadness_m destroyed, Bel_m\, Bel_j\, destroyed \rangle) \stackrel{def}{=} Bel_m\, Des_j \neg Bel_m\, destroyed$$

So the manager can apply this strategy, and as an effect he adopts the intention to receive compassion from agent j.

$$\mathit{Effect}(\langle \mathsf{SeekESupport}, m, Sadness_m destroyed, Bel_m\, Bel_j\, destroyed \rangle) \stackrel{def}{=} Intend_m\, SorryFor_{j,m} destroyed$$

To achieve this intention he will probably engage in a dialogue with j and use expressive dialog acts, but to receive compassion he must explain the cause of his sadness to his interlocutor.

8.5.2.3 Focus on and venting

This strategy is weaker than the previous one, since it has no additional condition, but it applies to all negative emotions. In any case, the manager can always use this strategy on an agent j attending his situation. The effect is that m adopts the intention to let j know about his emotion ($Intend_m\, Bel_j\, Sadness_m\, destroyed$). As for SeekESupport he can communicate his sadness verbally (engaging in dialogue to tell that he is sad and to explain his problem). However since he is not looking for compassion, he does not need to explain the cause of his emotion, so non-verbal venting of emotion (crying) would be sufficient.

8.5.2.4 Acceptation and resignation

The manager can also accept the situation and resign himself, through abandoning his contradicted desire. This strategy has no additional condition. As a result of its application, the manager drops his contradicted desire, *viz.* he tries to get used to the destruction of his hotel in order not to consider it undesirable anymore.

$$\mathit{Effect}(\langle \mathsf{Resign}, m, Sadness_m\, destroyed\rangle) \stackrel{def}{=} \neg Des_i\, \neg destroyed$$

By this way, his sadness disappears.

Actually, it may take some time to abandon a desire and detach oneself from the situation, but as a simplification we formalize the effect of this strategy as being immediate. This strategy is difficult to apply in a situation where the threatened desire is quite an important one, and the manager would probably not choose such a strategy in this case. But as we said before, the effective choice of a strategy by the decision process is out of the scope of this work.

8.5.2.5 Positive reinterpretation

In the same initial situation, we consider that the manager believes that the destruction of his hotel is an opportunity to rebuild it ($Bel_m\, G(destroyed \rightarrow canRebuild)$), and that this is not undesirable for him ($\neg Des_m\, \neg canRebuild$). The additional condition for the manager to positively reinterpret the destruction of his hotel thus holds with $canRebuild$ as the means of the strategy.

$$AC(\langle \mathsf{PosReinterp}, m, Sadness_m\, destroyed, canRebuild\rangle) \stackrel{def}{=}$$
$$Bel_m\, G(destroyed \rightarrow canRebuild) \wedge \neg Des_m\, \neg canRebuild$$

As an effect of the execution of this coping action, the manager now considers it desirable to rebuild his hotel ($Des_m\, canRebuild$), for example because it will be more beautiful than before.

$$\mathit{Effect}(\langle \mathsf{PosReinterp}, m, Sadness_m\, destroyed, canRebuild\rangle) \stackrel{def}{=}$$
$$Des_m\, canRebuild$$

He then reinterprets the event through the light of this positive consequence and abandon his contradicted desire ($\neg Des_m\, \neg destroyed$)[6]. Actually, the positive emotion triggered by the possibility to rebuild a better hotel makes him forget his sadness about the destruction. As a consequence he will probably engage in actions to rebuild his hotel.

[6]Thus this strategy is a special kind of acceptation.

8.5.2.6 Mental disengagement

In the same initial situation, we consider that the manager likes running because it makes him feel good ($Bel_m\ After_{m:run}\ feelsGood \wedge Des_m\ feelsGood$). The additional condition for mental disengagement of the destruction of the hotel through running thus holds.

$$AC(\langle \mathsf{MentalDisengage}, m, Sadness_m\, destroyed, Done_{m:run}\ \top \rangle) \stackrel{def}{=} Bel_m\ After_{m:run}\ feelsGood \wedge Des_m\ feelsGood$$

Actually, mental disengagement is an attempt to perform another satisfying action, with the aim that it triggers positive emotions that divert the individual from the current negative one. As a result of the strategy, the manager will adopt the intention to run.

$$\mathit{Effect}(\langle \mathsf{MentalDisengage}, m, Sadness_m\, destroyed, Done_{m:run}\ \top \rangle) \stackrel{def}{=} Intend_m\ Happens_{m:run}\ \top$$

Finally he will probably perform this action if all its conditions are true.

8.6 Discussion of other formalizations of coping

Meyer (2006, *cf.* Section 2.6.2) describes the triggering of four emotions relative to the agent's plans and goals, and the influence of the agent's emotions on its actions, following Oatley and Jenkins' theory of emotions (1996, *cf.* Section 1.2.1.1). He makes a correspondence between each emotion and one kind of action tendency (in the sense of Frijda (1986), for example anger induces aggression). Our approach differs from Meyer's one in several points. First, we do not represent the same phenomena. Following Lazarus, action tendencies are innately programmed reflexes, and thus unconscious, whereas coping strategies are conscious efforts to adapt to one's emotion. Moreover, each emotion corresponds to exactly one action tendency, while coping strategies can apply to several emotions, depending on the particular context. Second, our approach differs in its style, since Meyer provides the syntax and semantics of a programming language, aiming at immediately implementing these agents, whereas we are mainly interested in axiomatizing these agents' coping strategies in a BDI logic.

Gratch and Marsella (2004a, *cf.* Section 2.3.5) do not propose a logical account of psychological emotional processes but an implementation in the EMA agent. This agent's mental state is represented by a complex structure inspired from planning: the Causal Interpretation. The appraisal process triggers several emotions,

and the most intense one provides a coping opportunity. The different strategies are then assessed w.r.t. their coping potential, and the agent chooses and applies his preferred one. The implemented strategies are also inspired from the COPE model (Carver, Scheier, and Weintraub, 1989): planning, positive reinterpretation, acceptance, denial, mental disengagement, shift blame. Their effect on the Causal Interpretation is mainly expressed in terms of intention dropping or modification of utility or probability values. The authors regret to have a too direct link between appraisal and coping, while psychology underlines the complexity of this link. Yet in our work we do not handle this link at all. Finally the EMA agent is a functional implementation (Gratch and Marsella, 2004b), but it builds on a complex mental structure that the authors believe to be needed to represent the essential concepts involved in the description of appraisal and coping. On the contrary we believe that our logic of belief, intention, desire, probability, action and time also allows to represent all these concepts, in a more standard way since many agents build on a BDI architecture. Though such a logic makes it difficult to manage intensity degrees and to express causality (*cf.* Chapter 4).

Elliott's Affective Reasoner (1992, *cf.* Section 2.3.2) is a collection of programs simulating the emotional behaviour of humans. The agent can feel twenty-four emotions computed according to the OCC typology. Then, he disposes of a database containing several types of actions (*e.g.* somatic, behavioural, communicative, evaluative) among which he chooses depending on his personality and emotions. Some of these actions appear to be coping strategies (*e.g.* suppression, repression, communicating) while others are unconscious physiological manifestations of emotions (somatic actions), unintentional effects of emotions (obsessive attentional focus), action tendencies (behavioural responses), or reappraisals. Elliott thus implements some coping actions in his Affective Reasoner but he seems not to distinguish them from other types of behaviours, despite of psychological evidence supporting this distinction (Lazarus and Folkman, 1984). In our work, we only account for coping actions, as defined by Lazarus, and listed in the COPE model. Moreover, we do not account for the selection of a coping action because we believe that the correspondence is far more complex than a one-to-one pairing with the agent's emotion and personality. Finally, an implementation like the Affective Reasoner imposes to make some concessions, in particular to be less faithful to psychology. On the contrary, we tried to propose a psychologically sound logical account of coping strategies but our logic may be very complex (due to revision actions in particular) and difficult to implement.

8.7 Conclusion

We started this chapter by noticing that the close relationship between appraisal and coping, the two parts of the human emotional process, while being endorsed by the psychological literature, is rather neglected in virtual agents. Indeed, many researchers are interested in the triggering of emotions by an appraisal process, but very few manage their subsequent influence on their agent's behaviour. For example Meyer considers this influence through a formalization of the action tendencies, which is quite different from the coping process. Gratch and Marsella may have been the first ones to integrate appraisal and coping in an agent. To do so, they introduced a complex representation of the agent's mental state, that they believe to be necessary to express all the concepts needed to describe the emotional process. On the contrary we showed that BDI logics are expressive enough to describe emotional processes, *viz*. both appraisal (*cf*. Chapter 4) and coping (in this chapter). We thus believe that our logical framework allows to formalize the *coping* process in a simpler way than Gratch and Marsella's Causal Interpretation.

We proposed to represent coping strategies as actions, whose conditions and effects are expressed in terms of the agent's mental attitudes, and represented in a BDI logic. We believe that this well-known framework offers interesting properties, mainly its reusability in a large amount of existing BDI agents. The application domain for such agents mainly consists in designing human-like characters for virtual worlds: a plausible emotional model can increase their believability and thus improve the user's immersion in the virtual world. The ability to reason about another agent's emotions also opens applications in Ambient Intelligence, where such emotional agents could detect a human's emotions and help him to cope with them by proposing or executing coping strategies (*cf*. Chapter 6).

We would like to mention some shortcomings of our model. First, we do not formalize all coping strategies, but only those we believe to be most interesting for intelligent agents, and only concerning event-based emotions. Second, we do not manage the intensity and dynamics of emotions, a too big problem to be solved here, and thus we assume that the execution of a coping strategy simply makes the emotion disappear, instead of making its intensity decrease. Third, we only sketched the formal deductions in this chapter, and did not fully work out the belief change and belief preservation mechanisms at work. In particular, coping necessitates to abandon our preservation axioms for desires.

This preliminary account has been recently published in a slightly updated version (Adam and Longin, 2007). Now our short-term prospects consist in extending this model to account for agent-based emotions of the OCC typology. In a longer-term prospect we envisage to implement this model in an embodied conversational agent, in order to simulate some dialogic behaviours that are often observed in human-human interactions, but are not captured by actual models of dialogue (for example, why do people change subject suddenly, or refuse to answer a question or to believe obviousness). Indeed, we believe that such kinds of irrational behaviours follow from the use of coping strategies.

Conclusion

It is difficult to say what is impossible, for the
dream of yesterday is the hope of today
and the reality of tomorrow.
(Robert H. Goddard)

This thesis relates a multi-disciplinary project, starting from the understanding of the psychological definition of emotions, proceeding with their formalization and the deduction of their properties, and leading to their integration in implemented agents. This thesis brings multiple contributions. First it offers to the agent community a formal model of a rich set of emotions. Our formal definitions of emotions are intended to be as realistic as possible through different means: they are faithful to the psychological definitions and can capture the antecedent situations described in the original theory; they also allow to prove some intuitive properties of emotions; and finally human users were asked to assess the relevance of the emotions expressed by an agent using these definitions. This formal model of emotions thus enables researchers to integrate emotions in their agents without having to interpret and formalize a psychological theory by themselves. Indeed this is a difficult task that should be done once and for all and then reused when needed. That is why we also intended our formalisation to be as generic as possible, through using a well-known framework: BDI logics.

Second, this thesis highlights that BDI logics are a powerful tool to disambiguate complex concepts and reason about their properties. Indeed we were able to give formal thus unambiguous definitions of twenty emotions, and to prove theorems about their links with each other. BDI logics are often criticized and considered to be only a tool (or even a toy) for making formal but abstract proofs. Nevertheless these logics also allow to describe the architecture of an agent. We explored two different applications for our model (conversational agents and Ambient Intelligence) and even implemented it in a BDI agent.

However our applications remain incomplete since our formalism also is. Indeed, the human emotional process is made up of two components: appraisal,

leading to the triggering of emotions, and coping, leading to a subsequent adaptive modification of behaviour. In this thesis we only formalize appraisal so the triggered emotions have no influence on the agent's subsequent behaviour. That's why in the last chapter we provide an insight on our ongoing research about coping. Actually we try to formalize it in the same BDI framework. This line of research is still less explored than the triggering or the expression of emotions. So this thesis also opens interesting future prospects.

Our work can now be improved on several points. A first local enhancement relates to our covering of the OCC typology: we cannot formalize its object-based branch in our current logic. More globally, the limited expressivity of BDI logics (compared to natural language) entails that several aspects of emotions remain unexplored in our model. First, we do not compute the intensity of emotions. This prevents us from managing their decay over time or their blending. Second we do not describe their interaction with each other, *viz*. how several emotions combine to create mood, and how mood or other emotions can bias the triggering of a new emotion. Indeed this is linked with the respective intensities of the involved emotions. Third we could only approximate the link between an action and its effects, since no modal operator is currently completely axiomatized to account for this link. Moreover we did not investigate the concept of group emotions, that we believe would need the introduction of non standard operators to represent group mental attitudes. For instance we are currently trying to formalize group beliefs (Tuomela, 1992; Gaudou, Herzig, and Longin, 2007) and plan to do so with group ideals. Finally, the study of the complexity of our logic is out of the scope of this work.

More globally, we restricted our account to the cognitive aspect of emotions. However we said in the introduction that emotions are a multi-facet phenomenon: so we neglected the biological, physiological, or sociocultural aspects. One of these neglected aspects is the influence of culture and social norms on the expression or inhibition of the triggered emotions. Indeed there is a great variety of emotional expressivity across cultures. And we have only touched upon the problem of coping, that is yet crucial to be formalized in order to provide a complete account of human emotions.

Despite these limitations, our evaluation shows that the emotions that our model can simulate are perceived to be quite believable. Now we assume that believability is an aspect of a general property of interaction systems: they must inspire trust. Castelfranchi, Falcone, and Marzo (2006) show that users must trust in a system to use it. We are currently involved in a project about trust whose aim is to formalize this concept of trust into a BDI logic. This should allow to disambiguate this notion and automatically reason about it. This model would then be implemented in an agent and tested. Like this thesis, this is again a multi-disciplinary project

starting from a philosophical and sociological analysis of a concept and leading to its formalization and its use in an implemented system.

Finally this work is just a first step on the long path leading to the understanding of emotions. This path needs to be collectively explored by researchers from as various views as psychology, computer science, biology, sociology... This collaboration may seem difficult for now, but that is what it costs to understand what makes us human.

Bibliography

Adam, Carole (2006). PLEIAD : ProLog Emotionally Intelligent Agents Designer. un module de gestion des émotions d'un ACA. In Martin, J.-C. and C. Pélachaud, editors, *Workshop Francophone sur les Agents Conversationels Animés (WACA)*, Toulouse.

Adam, Carole and Fabrice Evrard (2005). Donner des émotions aux agents conversationnels. In Pesty, S. and J.-P. Sansonnet, editors, *WACA'01 - Premier Workshop francophone sur les Agents Conversationnels Animés*, Grenoble.

Adam, Carole, Fabrice Évrard, Benoit Gaudou, Andreas Herzig, and Dominique Longin (2006a). Modélisation logique d'agents rationnels pour l'intelligence ambiante. In *Actes des 14e Journées Francophones sur les Systèmes Multi-Agents (JFSMA 2006), Annecy, France, 18–20 octobre*, Vol. to appear. Hermès Science Publications.

Adam, Carole, Benoit Gaudou, Andreas Herzig, and Dominique Longin (2006b). A logical framework for an emotionally aware intelligent environment. In Augusto, Juan Carlos and Daniel Shapiro, editors, *1st ECAI Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI'06), Riva de Garda, Italy, August 29th*, `ftp://ftp.irit.fr/IRIT/LILAC/Adam_aitami2006.pdf`. IOS Press.

Adam, Carole, Benoit Gaudou, Andreas Herzig, and Dominique Longin (2006c). OCC's emotions: a formalization in a BDI logic. In Euzenat, Jérôme, editor, *Proc. of the Twelfth Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06), Varna, Bulgaria, september 13–15*, Vol. 4183 of *LNAI*, pp. 24–32. Springer-Verlag.

Adam, Carole, Andreas Herzig, and Dominique Longin (2007). PLEIAD: an emotional agent to assess the OCC typology. *RIA* Special issue: Modèles multi-agents pour des environnements complexes.

Adam, Carole and Dominique Longin (2007). Endowing emotional agents with coping strategies: from emotions to emotional behaviour. In *et al.*, C. Pelachaud, editor, *Intelligent Virtual Agents (IVA'07)*, p. to appear as a poster, Paris.

Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin (2004). An integrated theory of the mind. *Psychological Review* 111: 1036–1060.

Anderson, J.R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Anderson, J.R. (1993). *Rules of the Mind*. Lawrence Erlbaum Associates, NJ.

Anderson, J.R. and C. Lebiere (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

Anderson, J.R. and C. Lebiere (2003). *The Newell test for a theory of Mind*. Behavioral and Brain Sciences.

Aristotle (2003). *Politique*. Collection des Universités de France. Les Belles Lettres.

Arnold, M. B. (1950). An excitatory theory of emotion. In Reymert, L., editor, *Emotions and personality*. Academic Press, New York.

Arnold, Magda B. (1960). *Emotion and personality*. Columbia University Press, New York.

Austin, John L. (1962). *How To Do Things With Words*. Oxford University Press.

Averill, J. (1980). A constructionist view of emotion. In Plutchik, R. and H. Kellerman, editors, *Emotion: Theory, research, and experience*, Vol. 1, chapter 12. Academic Press, New York.

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

Bates, Joseph (1992). The nature of characters in interactive worlds and the Oz project. Technical report CMU-CS-92-200, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Becker, Christian, Stefan Kopp, and Ipke Wachsmuth (2004). Simulating the emotion dynamics of a multimodal conversational agent. In *ADS'04*. Springer LNCS.

Blackburn, P., M. Rijke, and Y. Venema (2001). *Modal logic*. Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Bower, G. H. (1991). Emotional mood and memory. *American Psychologist* 31: 129–148.

Bratman, Michael E. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, USA.

Broersen, Jan, Andreas Herzig, and Nicolas Troquard (2006). A STIT-extension of ATL, with applications in the epistemic and deontic domains. In Fisher, Michael and Wiebe Hoek, editors, *Proc. 10th Eur. Conf. on Logics in Artificial Intelligence (JELIA06), Liverpool, 13-15 September 2006*, Vol. 4160 of *LNAI*. Springer.

Burgess, John P. (2002). Basic tense logic. In Gabbay, Dov and Franz Guenthner, editors, *Handbook of Philosophical Logic*, Vol. 7, pp. 1–42. Kluwer Academic Publishers, 2nd edition.

Cannon, W.B. (1927). The james-lange theory of emotions: a critical examination and an alternative theory. *American Journal of Psychology* 39: 106–124.

Carofiglio, V. and F. de Rosis (2005). In favour of cognitive models of emotions. In *Workshop on Mind-Minding agents at AISB'05*.

Carver, C. S., M. F. Scheier, and J. K. Weintraub (1989). Assessing coping strategies: a theoretically based approach. *Journal of Personality Psychology* 56(2): 267–283.

Castelfranchi, C. and E. Lorini (2003). Cognitive anatomy and functions of expectations. In *IJCAI03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico.

Castelfranchi, Cristiano, Rino Falcone, and Francesca Marzo (2006). Being trusted in a social network: Trust as relational capital. *Lecture Notes on Artificial Intelligence (LNAI)* .

Castelfranchi, Cristiano and Fabio Paglieri (2007). The role of belief in goal dynamics: Prolegomena to a constructive theory of intentions. *Synthese* to appear.

Chellas, B. F. (1980). *Modal Logic: an Introduction*. Cambridge University Press.

Chellas, Brian (1992). Time and modality in the logic of agency. *Studia Logica* 51: 485–517.

Clore, G. L. and A. Ortony (2000). Cognition in emotion: Always, sometimes, or never? In Nadel, L., R. Lane, and G. L. Ahern, editors, *The Cognitive neuroscience of emotion*. Oxford University Press, New York.

Cohen, Philip R. and Hector J. Levesque (1990). Intention is choice with commitment. *Artificial Intelligence Journal* 42(2–3): 213–261.

Damasio, Antonio R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Pub Group.

Darwin, Charles R. (1872). *The expression of emotions in man and animals*. Murray, London.

Dastani, Mehdi and John-Jules Meyer (2006). Programming agents with emotions. In *Proc. 17th European Conf. on Artificial Intelligence (ECAI 2006), Trento, Italy, Aug. 28th–Sep. 1st*. IOS Press.

de Rosis, Fiorella, Catherine Pelachaud, Isabella Poggi, V. Carofiglio, and B. De Carolis (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.

Descartes, René (1649). *Les passions de l'âme*. Livre de Poche. LGF.

Durkheim, Emile (1961). *Moral Education: A Study in the Theory and Application of the Sociology of Education*. The Free Press, New York.

Ekman, P., W.V. Friesen, and J.C. Hager (2002). *Facial Action Coding System Invetigator's Guide*. A Human Face.

Ekman, Paul (1992a). Are there basic emotions? *Psychological review* 99(3): 550–553.

Ekman, Paul (1992b). An argument for basic emotions. *Cognition and Emotion* 6: 169–200.

Ekman, Paul (1999a). Basic emotions. In Dalgleish, T. and M. Power, editors, *Handbook of cognition and emotion*, chapter 3. John Wiley and Sons, Ltd, Sussex, UK.

Ekman, Paul (1999b). Facial expressions. In Dalgleish, T. and M. Power, editors, *Handbook of cognition and emotion*, chapter 16. John Wiley and Sons, Ltd, Sussex, UK.

Ekman, Paul and W. V. Friesen (1978). The facial action coding system. *Consulting Psychologist Press* .

El Jed, Mehdi (2006). Interactions sociales en univers virtuel: modèles pour une interaction située. Ph.D. diss., Université Paul Sabatier, Toulouse.

El Jed, Mehdi, Nico Pallamin, Julie Dugdale, and Bernard Pavard (2004). Modelling character emotion in an interactive virtual environment. In *AISB 2004 Convention: Motion, Emotion and Cognition*. The society for the study of Artificial Intelligence and the Simulation of Behaviour.

El Jed, Mehdi, Nico Pallamin, Lucila Morales, Carole Adam, and Bernard Pavard (2005). Interaction sociale multi-agents en univers virtuel: un aca chez les pompiers. `http://www.limsi.fr/aca/pages/05.11.15.journee.15nov05/ACA-morales-pallamin-eljed-pavard-adam.pdf`.

El Nasr, Magy Seif, John Yen, and Thomas R. Ioerger (2000). FLAME–Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems* 3(3): 219–257.

Elliott, Clark (1992). The Affective Reasoner : A process model of emotions in a multi-agent system. Ph.D. diss., Northwestern University, Illinois.

Elliott, Clark, Jeff Rickel, and James Lester (1999). Lifelike pedagogical agents and affective computing: An exploratory synthesis. *Lecture Notes in Computer Science* 1600: 195–211.

Fagin, Ronald, Joseph Y. Halpern, Moshe Y. Vardi, and Yoram Moses (1995). *Reasoning about knowledge*. MIT Press, Cambridge.

FIPA (Foundation for Intelligent Physical Agents) (2002). FIPA Communicative Act Library Specification. `http://www.fipa.org/repository/aclspecs.html`.

Folkman, S., R. S. Lazarus, C. Dunkel-Schetter, A. DeLongis, and R. J. Gruen (1986). Dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology* 50: 992–1003.

Forgas, J.P. (1995). Mood and judgment: The affect infusion model (aim). *Psychological Bulletin* 117: 39–66.

Frijda, N. H., A. Ortony, J. Sonnemans, and G. Clore (1992). The complexity of intensity: Issues concerning the structure of emotion intensity. *Review of personality and social psychology* 11: 60–89.

Frijda, N.H. (1986). *The emotions*. Cambridge University Press, New York.

Galati, D., K. R. Scherer, and P. Ricci-Bitti (1997). Voluntary facial expression of emotion: Comparing congenitally blind to normal sighted encoders. *Journal of Personality and Social Psychology* 73: 1363–1379.

Gaudou, Benoit, Andreas Herzig, and Dominique Longin (2007). Group belief and grounding in conversation. In Trognon, Alain, editor, *Language, cognition, interaction*. Presses Universitaire de Nancy, `http://www.univ-nancy2.fr/pun/`.

Gershenson, Carlos (1999). Modelling emotions with multidimensional logic. In *NAFIPS'99*. IEEE.

Gmytrasiewicz, Piotr J. and Christine L. Lisetti (2000). Using decision theory to formalize emotions in multi-agent systems. In *Fourth International Conference on Multi-Agent Systems*, pp. 391–392, Boston.

Gmytrasiewicz, Piotr J. and Christine L. Lisetti (2002). Emotions and personality in agent design and modeling. In Meyer, John-Jules Ch. and M. Tambe, editors, *Intelligent Agents VIII*, Vol. 2333 of *LNAI*, pp. 21–31, Berlin Heidelberg. Springer-Verlag.

Gratch, J. and S. Marsella (2004a). A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research* 5(4): 269–306.

Gratch, J. and S. Marsella (2004b). Evaluating the modeling and use of emotion in virtual humans. In *Proceedings of 3rd International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS-2004), New-York, USA, July 19th–23rd*, pp. 320–327. ACM.

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

Grice, H. Paul (1957). Meaning. *Philosophical Review* 66: 377–388.

Herzig, Andreas and Dominique Longin (2002). Sensing and revision in a modal logic of belief and action. In van Harmelen, F., editor, *Proc. of 15th European Conf. on Artificial Intelligence (ECAI 2002), Lyon, France, July 23–26*, pp. 307–311. IOS Press.

Herzig, Andreas and Dominique Longin (2004). C&L intention revisited. In Dubois, Didier, Chris Welty, and Mary-Anne Williams, editors, *Proc. 9th Int.*

Conf. on Principles of Knowledge Representation and Reasoning (KR 2004), Whistler, Canada, June 2–5, pp. 527–535. AAAI Press.

Herzig, Andreas and Nicolas Troquard (2006). Knowing How to Play: Uniform Choices in Logics of Agency. In Weiss, Gerhard and Peter Stone, editors, *5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan, 8–12 mai 2006*, pp. 209–216. ACM Press.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca.

Horty, John F. and Nuel Belnap (1995a). The deliberate Stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic* 24(6): 573–582.

Horty, John F. and Nuel Belnap (1995b). The deliberative stit: A study of action, omission, ability, and obligation. *Journal of Philosophical Logic* 24(6): 583–644.

Izard, C. E. (1993). Four systems for emotion activation: cognitive and noncognitive processes. *Psychological Review* 100(1): 68–90.

Izard, Caroll Ellis (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological review* 99(3): 561–565.

Izard, Carroll Ellis (1977). *Human emotions*. Springer.

Izard, Carroll Ellis (1984). Emotion-cognition relationships and human development. In lzard, C. E., J. Kagan, and R. Zajonc, editors, *Emotion, cognition, and behavior*, pp. 17–37. Cambridge University Press, New York.

James, W. (1884). What is an emotion? *Mind* 9: 188–205.

Jaques, Patricia A., Rosa M. Vicari, Sylvie Pesty, and Jean-Francois Bonneville (2004). Applying affective tactics for a better learning. In *In Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*. IOS Press.

Johnson-Laird, P. N. and K. Oatley (1992). Basic emotions: a cognitive science approach to function, folk theory and empirical study. *Cognition and Emotion,* 6: 201–223.

Jones, Andrew and José Carmo (2002). Deontic Logic and Contrary-to-duties. In Gabbay, Dov and Franz Guenthner, editors, *Handbook of Philosophical Logic*, Vol. 8, pp. 265–343. Kluwer Academic Publishers, 2nd edition.

Kripke, S. (1963). Semantical analysis of modal logic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 9: 67–96.

Lang, Jérôme, Leendert W. N. Van Der Torre, and Emil Weydert (2002). Utilitarian desires. *Journal of Autonomous Agents and Multi-Agent Systems* 5: 329–363.

Lang, P.J. (1994). The varieties of emotional experience: A meditation of James and Lange theory. *Psychological Review* 101: 211–221.

Lange, C.G. and W. James (1967). *The emotions*. Hafner, New York.

Larsen, Jeff T., A. Peter McGraw, Barbara A. Mellers, and John T. Cacioppo (2004). The agony of victory and thrill of defeat. mixed emotional reactions to disappointing wins and relieving losses. *Psychological science* 15(5): 325–330.

Laverny, Noel and Jérôme Lang (2004). From knowledge-based programs to graded belief-based programs. part i: On-line reasoning. In *16th European Conference on Artificial Intelligence(ECAI'04)*, pp. 368–372, Valencia, Spain. IOS Press.

Laverny, Noel and Jérôme Lang (2005a). From knowledge-based programs to graded belief-based programs. part ii: off-line reasoning. In *9th International Joint Conference on Artificiel Intelligence (IJCAI'05)*, pp. 497–502, Edinburgh, Scotland. Gallus.

Laverny, Noel and Jérôme Lang (2005b). From knowledge-based programs to graded belief-based programs, Part II: off-line reasoning . In *Proc. of the 9th International Joint Conference on Artificiel Intelligence (IJCAI'05) , Edinburgh, Scotland, 31/07/05-05/08/05*, pp. 497–502. Gallus.

Lazarus, R. S. (1984a). Thoughts on the relations between emotion and cognition. In Scherer, K. R. and P. Ekman, editors, *Approaches to emotion*, pp. 247–257. NJ: LawrenceErlbaum Associates Inc, Hillsdate.

Lazarus, Richard S. (1984b). On the primacy of cognition. *American Psychologist* 39(2): 124–129.

Lazarus, Richard S. (1991). *Emotion and Adaptation*. Oxford University Press.

Lazarus, Richard S. and Susan Folkman (1984). *Stress, Appraisal, and Coping*. Springer Publishing Company.

Lazarus, R.S. (1966). *Psychological Stress and the Coping Process*. McGraw-Hill, New York.

Lebiere, C. and J. R. Anderson (1993). A connectionist implementation of the act-r production system. In *Fifteenth Annual Conference of the Cognitive Science Society*, pp. 635–640.

Lester, James, Brian Stone, and Gary Stelling (1999). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction* 9: 1–44.

Lester, James C., Sharolyn A. Converse, Brian A. Stone, Susan E. Kahler, and S. Todd Barlow (1997). Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation. In du Boulay, B. and R. Mizoguchi, editors, *Artificial Intelligence in Education: Knowledge and Media in Learning Systems*. IOS Press.

Leventhal, H. and K. R. Scherer (1987). The relationship of emotion and cognition: A functional approach to a semantic controversy. *Cognition and Emotion* 1: 3–28.

Lorini, Emiliano, Andreas Herzig, and Cristiano Castelfranchi (2006). Introducing attempt in a modal logic of intentional action. In Fisher, Michael and Wiebe Hoek, editors, *European Conf. on Logic in AI (JELIA'06)*, LNAI, pp. 1–13, Liverpool (UK). Springer-Verlag.

Marsella, S. and J. Gratch (2003). Modeling coping behavior in virtual humans: don't worry, be happy. In *Proceedings of 2nd International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS-2003), Melbourne, Australia, July 14th–18th*, pp. 313–320. ACM.

Meyer, John Jules (2004). Reasoning about emotional agents. In de Mántaras, R. López and L. Saitta, editors, *16th European Conf. on Artif. Intell. (ECAI)*, pp. 129–133.

Meyer, John-Jules Ch. (2006). Reasoning about emotional agents. *International Journal of Intelligent Systems* 21(6): 601–619.

Minsky, Marvin L. (1988). *The society of mind*. Simon and Schuster.

Moffat, D., N. H. Frijda, and R. H. Phaf (1993). Analysis of a model of emotions. In Sloman, A., D. Hogg, G. Humphreys, A. Ramsay, and D. Partridge, editors, *Prospects for Artificial Intelligence: Proc. of AISB-93*, pp. 219–228, Amsterdam. IOS Press.

Nilsson, N. (2001). Teleo-reactive programs and the triple-tower architecture. *Electronic Transactions on Artificial Intelligence* 5: 99–110.

Oatley, K. and P. N. Johnson-Laird (1987). Towards a cognitive theory of emotions. *Cognition and Emotion* 1: 29–50.

Oatley, K. and P.N. Johnson-Laird (1996). The communicative theory of emotions: empirical tests, mental models, and implications for social interaction. In Martin, L.L. and A. Tesser, editors, *Striving and feeling: Interactions among goals, affect, and self-regulation*, pp. 363–393. Erlbaum, Mahwah, NJ.

Oatley, Keith (1992). *Best Laid Schemes: The Psychology of Emotions*. Cambridge University Press.

Oatley, Keith and Jennifer M. Jenkins (1996). *Understanding emotions*. Blackwell publishing.

Ochs, Magali, R. Niewiadomski, Catherine Pelachaud, and David Sadek (2005). Intelligent expressions of emotions. In *1st International Conference on Affective Computing and Intelligent Interaction ACII*, China.

Ochs, Magalie, Catherine Pélachaud, and David Sadek (2006). Les conditions de déclenchement des émotions d'un agent conversationnel empathique. In *WACA'2006*.

Ochs, Magalie, David Sadek, and Catherine Pelachaud (2007). Vers un modèle formel des émotions d'un agent rationnel dialoguant empathique. In *MFI'07*, p. to appear.

Ortony, A. and T.J. Turner (1990). What's basic about basic emotions? *Psychological review* 97(3): 315–331.

Ortony, Andrew (2003). On making believable emotional agents believable. In et al., R. Trappl, editor, *Emotions in Humans and Artifacts*, pp. 189–212. MIT Press.

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

Paiva, A., J. Dias, D. Sobral, and R. Aylett (2004). Caring for agents and agents that care: building empathic relations with synthetic agents. In *Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, New York.

Paiva, Ana and Isabel Machado (1998). Vincent: an autonomous pedagogical agent for on-the-job training. In Shute, V., editor, *Intelligent Tutoring Systems*. Springer-Verlag.

Paiva, Ana, Isabel Machado, and C. Martinho (1999). Enriching pedagogical agents with emotional behaviour: The case of vincent. In Johnson, Lewis, editor, *AIED Workshop on Instructional Uses of Synthetic Characters*.

Panksepp, J. (1992). A critical role for "affective neuroscience" in resolving what is basic about basic emotions. *Psychological review* 99(3): 554–560.

Pelachaud, C., V. Carofiglio, B. D. Carolis, F. D. Rosis, and I. Poggi (2002). Embodied contextual agent in information delivering application. In *Proceedings of First International Joint Conference on Autonomous Agents & Multi-Agent Systems (AAMAS-2002), Bologne*. ACM.

Pelachaud, Catherine and M. Bilvi (2003). Computational model of believable conversational agents. *Communication in Multi-Agent Systems* Background, current trnds and future(2650): 300–317.

Picard, R. W. (1997). Does HAL cry digital tears? emotions and computers. In G., Stork D., editor, *HAL's legacy*, chapter 13. MIT Press, Cambridge, MA.

Picard, R. W., E. Vyzas, and J. Healey (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Transactions Pattern Analysis and Machine Intelligence* 23(10).

Picard, Rosalind W. (1997). *Affective Computing*. MIT Press, Cambridge, MA.

Picard, Rosalind W. (1999). Affective computing for hci. In *HCI'99*, pp. 829–833.

Picard, Rosalind W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies* 59(1-2): 55–64.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In Plutchik, R. and H. Kellerman, editors, *Emotion: theory, research, and experiences*, Vol. 1: Theories of emotion, pp. 3–33. Academic.

Poggi, Isabella and Catherine Pelachaud (2000). Performative facial expressions in animated faces. In Cassell, J., J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pp. 155–188. MIT Press, Cambridge.

Prendinger, H. and M. Ishizuka (2001). Appraisal and filter programs for affective communication. In *AAAI Fall Symposium on Emotional and Intelligent*.

Prendinger, Helmut and Mitsuru Ishizuka (2005). Human physiology as a basis for designing and evaluating affective communication with life-like characters. *IEICE Transactions on Information and Systems* E88-D(11): 2453–2460.

Rao, Anand S. and Michael P. Georgeff (1991). Modeling rational agents within a BDI-architecture. In Allen, J. A., R. Fikes, and E. Sandewall, editors,

Proc. Second Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'91), pp. 473–484. Morgan Kaufmann Publishers.

Rao, Anand S. and Michael P. Georgeff (1992). An abstract architecture for rational agents. In Nebel, Bernhard, Charles Rich, and William Swartout, editors, *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pp. 439–449. Morgan Kaufmann Publishers.

Reilly, Neal (1996). Believable social and emotional agents. Ph.D. diss., School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Rickel, J. and W. Lewis Johnson (1997). Steve: an animated pedagogical agent for procedural training in virtual environements. In *Animated Interface Agents: making them intelligent*, pp. 71–76.

Roseman, I. J. (1984). Cognitive determinants of emotion: A structural theory. *Review of personality and social psychology* 5: 11–36.

Rousseau, Daniel and Barbara Hayes-Roth (1998). A social-psychological model for synthetic actors. In Sycara, Katia P. and Michael Wooldridge, editors, *2nd International Conference on Autonomous Agents (Agents'98)*, pp. 165–172, New York. ACM Press.

Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology* 39: 1161–1178.

Russell, James A. (1997). How shall an emotion be called? In Plutchik, R. and H.R. Conte, editors, *Circumplex models of personality and emotions*, pp. 205–220. American Psychological Association, Washington, DC.

Sadek, M. D. (1992). A study in the logic of intention. In Nebel, Bernhard, Charles Rich, and William Swartout, editors, *Proc. Third Int. Conf. on Principles of Knowledge Representation and Reasoning (KR'92)*, pp. 462–473. Morgan Kaufmann Publishers.

Sahlqvist, H. (1975). Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., editor, *Proc. 3rd Scandinavian Logic Symposium*, Vol. 82 of *Studies in Logic*.

Schacter, S. and J.E. Singer (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review* 69: 379–399.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. In Scherer, Klaus Rainer, Angelika Schorr, and Tom Johnstone, editors, *Appraisal Processes in Emotion : Theory, Methods, Research*, pp. 92–120. Oxford University Press, New York.

Scherer, Klaus Rainer, Angelika Schorr, and Tom Johnstone (2001). *Appraisal processes in emotion: theory, methods, research*. Oxford University Press.

Scherer, K.R. (1984). Emotion as a multicomponent process: a model and some cross-cultural data. *Review of personality and social psychology* 5: 37–63.

Scherer, K.R. (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. and J. Sangsue (1995). Le système mental en tant que composant de l'émotion. In *XXVe Journées d'Études de l'Association de Psychologie Scientifique de Langue Française (APSLF)*, Coimbra, Portugal.

Searle, J. R. and D. Vanderveken (1985). *Foundation of illocutionary logic*. Cambridge University Press.

Searle, John R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York.

Searle, John R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Sloman, Aaron (2001). Varieties of affect and the cogaff architecture schema. In Johnson, C., editor, *Symposium on Emotion, Cognition, and Affective Computing AISB'01 Convention*, pp. 39–48, York.

Smith, C. A. and P. C. Ellsworth (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology* 48: 813–838.

Spinoza (1994). *L'éthique*. Folio. Gallimard.

Staller, Alexander and Paolo Petta (2001). Introducing emotions into the computational study of social norms: a first evaluation. *Journal of artificial societies and social simulation* 4(1).

Tomkins, Silvan S. (1962). *Affect, imagery, consciousness*, Vol. 1: The positive affects. Springer, New York.

Tomkins, Silvan S. (1963). *Affect, imagery, consciousness*, Vol. 2: The negative affects. Springer, New York.

Tomkins, Silvan S. (1980). Affect as amplification: Some modifications in theory. In Plutchik, R. and H. Kellerman, editors, *Emotion: Theory, research, and experience*, Vol. 1: Theories of emotion, pp. 141–164. Academic Press, New York.

Tomkins, Silvan S. and C. Izard (1965). *Affect, cognition, and personality: Empirical studies*. Springer, New York.

Troquard, Nicolas, Robert Trypuz, and Laure Vieu (2006). Towards an ontology of agency and action : From stit to ontostit+. In Bennett, Brandon and Christiane Fellbaum, editors, *International Conference on Formal Ontology in Information Systems*, Vol. 150 of *Frontiers in Artificial Intelligence and Applications*, pp. 179–190, Baltimore, Maryland, USA. IOS Press.

Tuomela, Raimo (1992). Group beliefs. *Synthese* 91: 285–318.

Valins, Stuart C. (1966). Cognitive effects of false heart-rate feedback. *Journal of Personality and Social Psychology* 4: 400–408.

van der Hoek, W., B. Linder, and J-J. Ch. Meyer (1998). An integrated modal approach to rational agents. In Wooldridge, M. and A. Rao, editors, *Foundations of Rational Agency*, Vol. 14 of *Applied Logic*, pp. 133–168. Kluwer, Dordrecht.

Velàsquez, Juan D. (1997). Modeling emotions and other motivations in synthetic agents. In *AAAI'97*.

Velàsquez, Juan D. and Pattie Maes (1997). Cathexis: a computational model of emotions. In *First International conference on autonomous agents (AGENTS'97)*, pp. 518–519. ACM Press.

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

Wooldridge, Michael (2000). *Reasoning about rational agents*. MIT Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist* 35: 151–175.

Zajonc, R. B. (1984). On primacy of affect. In Scherer, K. R. and P. Ekman, editors, *Approaches to emotion*, pp. 270–259. NJ: Lawrence Erlbaum Associates Inc, Hillsdate.

Part IV

Appendix

Appendix A

Summary of axiomatics

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

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

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

$$(\Box A \wedge \Box B) \to \Box(A \wedge B) \qquad \text{(C-}\Box\text{)}$$

$$(\Box A \wedge \Diamond B) \to \Diamond(A \wedge B)$$

$$\frac{A \to B}{\Diamond A \to \Diamond B} \qquad \text{(RK-}\Diamond\text{)}$$

$$\mathit{Happens}_\alpha\,\varphi \to \mathit{After}_\beta\,\varphi \qquad \text{(CD-HA)}$$

$$\mathit{Done}_\alpha\,\varphi \to \mathit{Before}_\beta\,\varphi \qquad \text{(CD-DB)}$$

$$\varphi \to \mathit{After}_\alpha\,\mathit{Done}_\alpha\,\varphi \qquad \text{(CONV-AD)}$$

$$\varphi \to \mathit{Before}_\alpha\,\mathit{Happens}_\alpha\,\varphi \qquad \text{(CONV-BH)}$$

$$\mathit{Bel}_i\,\varphi \to \neg\mathit{Bel}_i\,\neg\varphi \qquad \text{(D-}\mathit{Bel}_i\text{)}$$

$$\mathit{Bel}_i\,\varphi \to \mathit{Bel}_i\,\mathit{Bel}_i\,\varphi \qquad \text{(4-}\mathit{Bel}_i\text{)}$$

$$\neg\mathit{Bel}_i\,\varphi \to \mathit{Bel}_i\,\neg\mathit{Bel}_i\,\varphi \qquad \text{(5-}\mathit{Bel}_i\text{)}$$

$$G\varphi \rightarrow \varphi \qquad \text{(T-}G\text{)}$$
$$(F\varphi \wedge F\psi) \rightarrow F(\varphi \wedge F\psi) \vee F(\psi \wedge F\varphi) \qquad \text{(3-}F\text{)}$$
$$G\varphi \rightarrow GG\varphi \qquad \text{(4-}G\text{)}$$
$$H\varphi \rightarrow \varphi \qquad \text{(T-}H\text{)}$$
$$(P\varphi \wedge P\psi) \rightarrow P(\varphi \wedge P\psi) \vee P(\psi \wedge P\varphi) \qquad \text{(3-}P\text{)}$$
$$H\varphi \rightarrow HH\varphi \qquad \text{(4-}H\text{)}$$
$$\varphi \rightarrow GP\varphi \qquad \text{(CONV-GP)}$$
$$\varphi \rightarrow HF\varphi \qquad \text{(CONV-HF)}$$

$$\frac{\varphi \rightarrow \psi}{Prob_i\,\varphi \rightarrow Prob_i\,\psi} \qquad \text{(RM-}Prob_i\text{)}$$
$$\frac{\varphi}{Prob_i\,\varphi} \qquad \text{(RN-}Prob_i\text{)}$$
$$Prob_i\,\varphi \rightarrow \neg Prob_i\,\neg\varphi \qquad \text{(D-}Prob_i\text{)}$$

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

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

$$Prob_i\,\varphi \rightarrow Bel_i\,Prob_i\,\varphi \qquad \text{(4-MIX1)}$$
$$\neg Prob_i\,\varphi \rightarrow Bel_i\,\neg Prob_i\,\varphi \qquad \text{(5-MIX1)}$$
$$Des_i\,\varphi \rightarrow Bel_i\,Des_i\,\varphi \qquad \text{(4-MIX2)}$$
$$\neg Des_i\,\varphi \rightarrow Bel_i\,\neg Des_i\,\varphi \qquad \text{(5-MIX2)}$$

$$Done_\alpha\,\top \rightarrow Bel_i\,Done_\alpha\,\top \qquad \text{(4-MIX3)}$$
$$\neg Done_\alpha\,\top \rightarrow Bel_i\,\neg Done_\alpha\,\top \qquad \text{(5-MIX3)}$$

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

$$Bel_i\,\varphi \rightarrow Prob_i\,\varphi$$
$$Prob_i\,\varphi \rightarrow \neg Bel_i\,\neg\varphi$$

$$G\varphi \rightarrow After_\alpha\,\varphi \qquad \text{(GA-MIX)}$$
$$H\varphi \rightarrow Before_\alpha\,\varphi \qquad \text{(HB-MIX)}$$

$$Des_i\,\varphi \rightarrow GDes_i\,\varphi \qquad \text{(Pers-}Des_i\text{)}$$
$$\neg Des_i\,\varphi \rightarrow G\neg Des_i\,\varphi \qquad \text{(Pers-}\neg Des_i\text{)}$$

$$Idl_i\,\varphi \rightarrow GIdl_i\,\varphi \qquad \text{(Pers-}Idl_i\text{)}$$
$$\neg Idl_i\,\varphi \rightarrow G\neg Idl_i\,\varphi \qquad \text{(Pers-}\neg Idl_i\text{)}$$

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

Appendix B

Summary of formal definitions of emotions

Formal definition (Well-being emotions).

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

Formal definition (Prospect-based emotions).

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

Formal definition (Confirmation emotions).

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

Formal definition (Disconfirmation emotions).

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

Formal definition (Good-will fortunes-of-others emotions).

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

Formal definition (Ill-will fortunes-of-others emotions).

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

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

Formal definition (Self-agent attribution emotions).

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

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

Formal definition (Other agent attribution emotions).

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

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

Formal definition (Composed emotions).

$$Gratification_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} Pride_i\,(i{:}\alpha, \varphi) \wedge Joy_i\,\varphi$$

$$Remorse_i\,(i{:}\alpha, \varphi) \stackrel{def}{=} Shame_i\,(i{:}\alpha, \varphi) \wedge Distress_i\,\varphi$$

$$Gratitude_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Admiration_{i,j}(j{:}\alpha, \varphi) \wedge Joy_i\,\varphi$$

$$Anger_{i,j}(j{:}\alpha, \varphi) \stackrel{def}{=} Reproach_{i,j}(j{:}\alpha, \varphi) \wedge Distress_i\,\varphi$$

Appendix C

Résumé de la thèse en français

Introduction

Nous humains ressentons des émotions tous les jours, presque à chaque minute, pas toujours consciemment. Nos émotions jouent un rôle crucial dans nos vies et nos relations avec les autres. Elles peuvent être manipulées par des publicités qui veulent nous faire acheter, par des hommes politiques qui veulent nous faire voter pour eux, ou par des personnes voulant nous séduire. Elles peuvent nous faire agir impulsivement, ce que nous regrettons ensuite. Elles peuvent nous compliquer la vie et interférer avec notre travail en nous perturbant. Nous cherchons souvent à les cacher, pour respecter des normes sociales ou pour ne pas révéler ce que nous pensons. Finalement, nous croyons en général devoir contrôler nos émotions pour être rationnels ou pour être performants. D'ailleurs, de plus en plus d'athlètes ont recours à des préparateurs mentaux qui leur apprennent à contrôler leurs émotions. En fait de plus en plus de gens font appel à ce genre de coaches, ou lisent des livres supposés les aider à contrôler les émotions indésirables.

Cependant il est d'autant plus difficile de contrôler ses émotions que nous ne les comprenons même pas. En fait nous ne savons même pas exactement ce qu'est une émotion, nous serions incapables d'en donner une définition claire. Et pourtant, beaucoup de sciences ont tenté d'établir une telle définition, mais la difficulté vient du fait que les émotions sont un phénomène complexe associant différents types de réactions interdépendantes, relevant de différentes sciences : la physiologie, la psychologie, la biologie... Ainsi chacune de ces sciences ne peut donner qu'une définition partielle d'un aspect des émotions.

La recherche d'une définition des émotions n'a longtemps intéressé que les philosophes, qui les voyaient comme des instincts nuisibles (passions, (Descartes, 1649)), opposés à la raison (Spinoza, 1994), dont l'homme doit se débarrasser

(*catharsis*, (Aristotle, 2003)). Puis la biologie a montré que les émotions étaient des réflexes utiles pour la survie, hérités au cours de l'évolution (Darwin, 1872), et a proposé des classifications des émotions basiques. La physiologie a ensuite tenté d'expliquer les mécanismes émotionnels, et plusieurs théories ont émergé, en désaccord sur le centre émotionnel qu'elles identifiaient (Lange and James, 1967; Cannon, 1927). La psychologie accepte le rôle adaptatif des émotions dans la vie humaine et propose diverses théories pour expliquer leur déclenchement d'un point de vue cognitif (Schacter and Singer, 1962; Arnold, 1950). La sociologie s'intéresse plus particulièrement à la construction sociale des émotion et à leur rôle dans le maintien de la cohésion d'un groupe (Averill, 1980; Durkheim, 1961). Cependant, malgré la profusion de théories soutenant le rôle essentiel des émotions pour les individus comme pour la société, l'informatique néglige toujours ce phénomène. Pour le développement d'agents rationnels, les émotions sont jugées trop complexes et non pertinentes.

Heureusement, les progrès des neurosciences (Damasio, 1994) leur ont apporté des preuves plus tangibles du rôle des émotions dans le comportement intelligent, la prise de décision, la planification, la communication sociale, et toutes ces capacités humaines supposées rationnelles : nous humains ne pouvons pas raisonner si nous ne ressentons pas d'émotions. Bates (1994) introduit alors le concept d'agents "crédibles" (*believable agents*), des agents qui donnent l'illusion d'être vivants, comme les personnages de dessins animés, et il montre que les émotions sont cruciales dans la création de tels agents. Enfin Picard (1997) est la première informaticienne à argumenter que les agents virtuels ont besoin d'être dotés d'émotions pas seulement pour être crédibles, mais aussi pour être vraiment intelligents et pour interagir de manière naturelle et amicale avec les humains. Elle introduit le concept d'*Affective Computing* et l'informatique commence alors à réellement s'intéresser aux émotions. La communauté agent commence à créer des agents émotionnels pour diverses applications : interfaces homme-machine intuitives, agents crédibles pour peupler les mondes virtuels, agents pédagogiques intelligents... Ces agents doivent non seulement exprimer des émotions, mais aussi percevoir et comprendre celles de l'utilisateur pour adapter leur comportement en conséquence. Bien sûr ce qu'on appelle émotions pour ces agents ne correspond pas exactement à ce que sont les émotions humaines. Les *émotions virtuelles* données aux agents virtuels sont plutôt des sortes d'étiquettes permettant de décrire un état particulier qui affecte leur comportement (Picard, 1997) : quand un agent est dans cet état il exprime un comportement cohérent avec l'émotion associée, même s'il ne "ressent" pas vraiment cette émotion comme nous la ressentons.

Ainsi la création d'agents émotionnels ne peut se faire que par une coopération entre l'informatique et la psychologie (Gratch and Marsella, 2005). L'informatique

doit absolument s'inspirer des théories psychologiques qui ont déjà essayé de décoder les émotions depuis des décennies. La plupart des créateurs d'agents, à la recherche d'une théorie compréhensible et adaptée à leurs besoins, se basent sur la typologie des émotions d'Ortony, Clore, and Collins (1988) qui était prévue pour être utilisée dans des applications d'Intelligence Artificielle. Mais chacun en procure ensuite sa propre formalisation ou implémentation, souvent partielle, spécialement conçue pour ses agents.

Au contraire nous pensons que comprendre et formaliser une théorie est un travail conséquent qui devrait être fait une fois pour toutes, puis réutilisé par tous les créateurs d'agents pour leur éviter de perdre du temps et de l'énergie à repartir de zéro. De plus les implémentations directes d'une théorie dans un agent sont spécifiques au domaine d'application de cet agent et ne sont pas réutilisables dans d'autres agents ou pour d'autres applications. Il est donc important de passer par une phase de formalisation qui propose un modèle générique réutilisable. Nous pensons que les logiques formelles permettent de créer de tels modèles.

En effet, les émotions sont un phénomène complexe dont la définition est souvent abstraite, ambiguë, subjective et non consensuelle. Les logiques formelles offrent un vocabulaire universel à la sémantique claire qui permet aux informaticiens de désambiguïser ces concepts. Elles permettent aussi de raisonner, de planifier et d'expliquer le comportement des agents. La formalisation logique d'un phénomène peut même révéler des problèmes non intuitifs. Les logiques BDI, *viz*. les logiques de la croyance (*Belief*), du désir (*Desire*) et de l'intention (*Intention*) (Cohen and Levesque (1990), Rao and Georgeff (1991, 1992), Sadek (1992), Herzig and Longin (2004), Wooldridge (2000)) se basent sur la philosophie du langage, de l'esprit et de l'action (*cf*. Bratman (1987), Searle (1969, 1983)). Elles proposent de modéliser les agents grâce à des concepts comme l'action et les *attitudes mentales* (croyances, buts, intentions, choix...). Ce cadre est communément utilisé dans la communauté agent et offre des propriétés intéressantes : grand pouvoir explicatif, vérifiabilité formelle, cadre théorique rigoureux et bien établi (à la fois du point de vue de la philosophie et de la logique formelle).

Par ailleurs, Searle (1983) soutient que les émotions sont des attitudes mentales particulières et peuvent être exprimées en termes de croyances et de désirs, ce qui encourage l'idée d'utiliser les logiques BDI pour les représenter. Pourtant, il y a encore peu de travaux au sujet des émotions dans la communauté des logiciens. Par exemple Meyer (2004) propose un langage basé sur les logiques BDI pour la description d'agents émotionnels, et Ochs et al. (2005) procure des définitions de quelques émotions en utilisant un formalisme BDI particulier, la théorie de l'interaction rationnelle de Sadek (1992). Mais ces modèles ne profitent pas de tous les avantages des logiques BDI. En particulier ils s'intéressent plus à l'implémentation de leur modèle dans un agent qu'aux possibilités de raison-

nement formel qu'il offre. De plus ils ne sont pas toujours très fidèles aux théories psychologiques sous-jacentes, et ne proposent qu'un jeu assez restreint d'émotions.

C'est pourquoi l'objectif de cette thèse est de fournir un modèle logique générique des émotions, et plus particulièrement de leur aspect cognitif au détriment des aspects biologiques ou physiologiques. Notre modèle se voudra fidèle à la psychologie, proposera un grand nombre d'émotions, et aura une axiomatique correcte et complète pour permettre de déduire des propriétés des émotions. Cependant notre modèle ne devra pas être réduit à cet usage théorique, nous l'exploiterons donc dans des applications pratiques et l'implémenterons dans un agent à architecture BDI. Cet agent devra alors exposer des émotions crédibles en réponse aux stimuli, ce que nous évaluerons.

Ce mémoire est structuré en trois parties. La première partie est consacrée à l'état de l'art, du point de vue de la psychologiqe puis de l'informatique. Nous introduirons un historique des principales théories psychologiques des émotions, avant d'expliquer quelle théorie nous choisissons de formaliser et pourquoi (Chapter 1). Nous exposerons alors quelques travaux qui ont déclenché l'intérêt de l'informatique pour les émotions, et décrirons quelques applications d'Intelligence Artificielle impliquant des agents émotionnels (Chapter 2).

La deuxième partie constitue le coeur de ce travail, *viz.* notre formalisation des émotions. Nous introduirons notre formalisme BDI particulier, donnerons sa sémantique et son axiomatique et prouverons sa correction et sa complétude (Chapter 3). Nous formaliserons alors vingt émotions en termes des opérateurs modaux ainsi introduits ; nous discuterons nos choix de formalisation ainsi que les différences entre plusieurs théories psychologiques décrivant la même émotion ou des émotions proches (Chapter 4). Une fois ces définitions acceptées, notre logique nous permettra de prouver quelques propriétés intuitives des émotions (Chapter 5). Cela soutient la justesse de nos définitions ainsi que la puissance des logiques BDI pour désambiguïser des concepts complexes.

Finalement la troisième partie exposera quelques applications concrètes et travaux en cours qui découlent de cette formalisation. En particulier nous avons mené quelques petites études de cas dans le domaine de l'Intelligence Ambiante et du dialogue machine-machine (Chapter 6). Nous décrirons ensuite l'implémentation de notre formalisme dans l'agent PLEIAD et son utilisation pour évaluer la crédibilité des émotions générées (Chapter 7). Finalement nous exposerons des résultats préliminaires au sujet de la formalisation dans la même logique du processus *coping*, qui décrit comment un agent pourrait adapter son comportement pour gérer ses émotions négatives (Chapter 8).

C.1 Les émotions en psychologie

Dans ce chapitre, nous commençons par présenter un historique de l'évolution des théories psychologiques. Nous distinguons alors deux grandes tendances dans ces théories. La première tendance en psychologie des émotions a été de chercher à identifier un nombre limité d'émotions, et d'essayer de les classer. C'est seulement plus tard que des théories ont été élaborées pour tenter d'expliquer le fonctionnement des émotions. Les théories appartenant à ces deux tendances ne répondent pas aux mêmes questions ; en fait la plupart des théories s'intéressent principalement à une seule de ces grandes questions, et ne donnent pas de réponse explicite à la deuxième.

Parmi les théories qui s'intéressent plus particulièrement à la **classification** des émotions, on trouve deux tendances opposées : différents types d'approches considèrent (pour différentes raisons) qu'il existe un nombre limité d'émotions basiques ; au contraire quelques théories dites continues s'opposent à ce point de vue (*cf.* Section 1.2.1).

Les théories qui s'intéressent plutôt au **fonctionnement** des émotions ne donnent pas forcément de réponse à cette question de l'existence d'émotions basiques. Elles cherchent à répondre à plusieurs autres questions : comment les émotions sont-elles déclenchées ? quelle est leur influence sur le comportement ? en particulier est-ce que les émotions déclenchent ou sont déclenchées par les modifications physiologiques ? est-ce que la cognition joue un rôle dans le déclenchement des émotions ? Les théories physiologiques (*cf.* Section 1.2.2) et cognitives (*cf.* Section 1.2.3) ont des points de vue opposés sur cette dernière question.

Dans la suite du chapitre nous décrivons quelques théories représentant chacun de ces courants de recherche.

C.1.1 Approches intéressées par la classification des émotions

Deux grandes tendances s'opposent au sein de ce courant de recherche sur la classification des émotions. La première suppose qu'il existe un nombre limité d'émotions basiques, alors que la deuxième considère les émotions comme une fonction continue de deux ou trois dimensions. En fait Ekman (1999a) repère trois définitions différentes de la notion d'émotion basique :

- des émotions adaptatives en relation avec une théorie évolutionniste : les émotions basiques sont celles qui ont été conçues au cours de l'évolution pour répondre à des problèmes de survie spécifiques, et sont maintenant des réponses innées programmées dans les individus (*e.g.* Darwin, Ekman, Oatley and Johnson-Laird);

- des émotions discrètes qui diffèrent les unes des autres de manière importante, mais pas forcément en rapport avec une théorie évolutionniste (*e.g.* Tomkins, Izard) ; c'est à cette hypothèse que s'opposent les théories continues qui considèrent que les émotions sont toutes similaires en essence et ne diffèrent que par les valeurs de quelques dimensions comme l'intensité ou la valence (*e.g.* Lang, Russell);

- des "building blocks" : ceci est la point de vue de Plutchik, qui compare les émotions basiques aux couleurs primaires, pouvant se combiner pour construire les émotions complexes ; ce point de vue est en contradiction avec la notion évolutionniste d'émotion basique selon laquelle deux émotions basiques ne peuvent pas coexister en même temps.

À une époque où les émotions n'intéressaient que la philosophie, qui les considéraient comme irrationnelle, Darwin (1872) a proposé la première théorie moderne des émotions. Plusieurs chercheurs ont ensuite pris sa suite et proposé d'autres théories évolutionnistes. Par exemple Ekman (1992b; 1992a; 1999a; 1999b) se base sur les expressions faciales pour distinguer six émotions de base. Toutes ces théories partagent l'hypothèse que les humains ont hérité au cours de l'évolution d'un petit nombre d'émotions basiques à valeur adaptative et communicative, devenus des réflexes innés, sans composante cognitive. Sur ce point ces théories s'opposent donc aux théories cognitives que nous détaillerons plus tard. Les théories évolutionnistes supposent aussi toutes que ce sont les émotions qui déclenchent des modifications physiologiques à but adaptatif. Sur ce point elles diffèrent des autres théories discrètes. De plus les théories évolutionnistes diffèrent les unes des autres au sujet des émotions particulières qu'elles considèrent comme étant basiques.

D'autres théories s'accordent sur l'existence d'un nombre limité d'émotions basiques, sans supposer qu'elles auraient été héritées au cours de l'évolution. Ces théories discrètes s'opposent aux théories continues selon lesquelles les émotions sont des phénomènes tous similaires, ne variant que par la valeur de certaines dimensions. De plus Tomkins et Izard supposent que les émotions sont déclenchées par la perception des modifications physiologiques de l'individu, au contraire des théories évolutionnistes.

Finalement il existe de nombreuses théories discrètes, qui s'accordent toutes sur l'existence d'émotions basiques, mais qui n'ont pu s'accorder sur la définition même d'une émotion basique. Ainsi il n'existe encore aucun consensus sur la nature, le nombre et l'identité des émotions qui seraient basiques (*cf.* Table 1.12). Ce manque de consensus a rendu les chercheurs d'autres branches très sceptiques quant à l'existence d'émotions basiques, menant à de grands débats à ce sujet. Actuellement les points de vue diffèrent encore sur la structure de l'espace émotionnel. En même temps, d'autres types de théories se focalisent sur d'autres as-

pects des émotions sans répondre à cette question. Ce sont les théories plutôt intéressées par le déclenchement des émotions et leur fonctionnement.

C.1.2 Approches intéressées par le fonctionnement des émotions

Nous avons séparés ces approches en deux tendances : les théories physiologiques (Section 1.2.2) qui soulignent le rôle des modifications physiologiques dans le déclenchement des émotions, et les théories cognitives (Section 1.2.3) qui soulignent la primauté de la cognition dans le déclenchement des émotions.

Les approches physiologiques considèrent que l'activation physiologique est l'unique origine des émotions. Elles supposent que les stimuli déclenchent une activation physiologique (qui peut concerner les viscères, le centre vasomoteur, ou le thalamus, selon les théories) qui crée les émotions. Mais des expériences ultérieures ont montré que l'activation seule n'était pas suffisant pour différencier les émotions les unes des autres. Des chercheurs font alors l'hypothèse que cette activation doit être cognitivement interprétée pour déclencher une émotion particulière (Schacter and Singer, Valins). Ainsi apparaissent les premières théories cognitives. Puis Arnold introduit le concept d'évaluation cognitive ("appraisal"), un processus d'évaluation non pas de l'activation physiologique mais du stimulus, par rapport à divers critères. De nombreux chercheurs proposent alors des théories de l'évaluation cognitive comprenant des ensembles de critères différents (Frijda, Lazarus, Scherer, Ortony, Clore et Collins). Enfin les théories schématiques suggèrent que les émotions sont en fait activées quand leurs éléments cognitifs le sont (Leventhal). Le point commun de toutes ces théories cognitives est de souligner l'importance de la cognition dans la naissance des émotions, ce qui les oppose aux théories physiologiques ou évolutionnistes. Il semblerait cependant que ce désaccord provienne de l'utilisation différente du concept de cognition, et finalement les chercheurs semblent s'accorder sur le rôle de cognitions de plus ou moins bas niveau dans les émotions.

C.1.3 Choix d'une théorie à formaliser

Dans cet état de l'art nous avons présenté diverses théories des émotions qui divergent sur plusieurs grandes questions. Nous allons maintenant expliquer pourquoi nous choisissons de formaliser une théorie de l'évaluation cognitive plutôt qu'un autre type de théorie, et laquelle en particulier. Pour cela nous exposons ici les avantages des théories de l'*appraisal* tels qu'ils ont été identifiés par (Scherer, Schorr, and Johnstone, 2001).

Tout d'abord ces théories permettent une différenciation fine entre les émotions, en les faisant correspondre chacune à un motif d'évaluation particulier, c'est-

à-dire à une certaine configuration des critères d'évaluation. Cette médiation cognitive entre le stimulus et l'émotion est aussi la seule manière d'expliquer les différences inter-individuelles : dans la même situation deux individus peuvent ressentir des émotions très différente, ce qui résulte en fait de l'évaluation différente qu'ils font de la situation. Cette médiation rend aussi compte de la variabilité temporelle des émotions : un même individu peut évaluer différemment la même situation à différents moments. Une correspondance directe entre stimulus et émotion ne pourrait expliquer aucun de ces phénomènes pourtant établis. De plus les théories évolutionnistes prétendant que les émotions sont des réponses innées à des situations connues ne peuvent expliquer comment des émotions se produisent face à des situations inédites, inconnues. Enfin les théories de l'évaluation cognitive expliquent pourquoi les réponses émotionnelles sont appropriées au contexte, puisque c'est l'évaluation de ce contexte qui les déclenche.

C'est pourquoi dans cette thèse nous choisissons de formaliser une théorie de l'évaluation cognitive, et plus particulièrement la typologie OCC (Ortony, Clore, and Collins, 1988) qui a été conçue pour des applications en IA. Dans ce chapitre nous donnerons donc plus de détails sur cette théorie particulière afin de permettre au lecteur de comprendre la formalisation qui suivra.

C.1.4 Conclusion

Ce qu'il faut retenir de ce chapitre, c'est que malgré la profusion de théories s'attaquant à l'explication de divers aspects des émotions (principalement leur classification et leur fonctionnement), aucun consensus n'a encore été trouvé sur les nombreux problèmes qui se posent. Il est donc difficile pour les informaticiens de s'y retrouver. C'est pourquoi il nous paraît tellement essentiel que le travail de formalisation des théories psychologiques soit fait une fois pour toutes et puisse ensuite être réutilisé facilement par les chercheurs.

C.2 Les émotions en informatique

Dans ce chapitre nous nous intéressons à l'utilisation des théories psychologiques des émotions dans un domaine qui peut paraître complètement opposé : l'informatique. Après avoir négligé les émotions pendant des décennies, l'informatique commence à s'intéresser à leur potentiel pour des agents artificiels dans les années 90. Il est important de remarquer que ce qu'on appelle émotions pour un agent virtuel ne correspond pas exactement à ce qu'on appelle émotions chez un humain. Comme le dit Picard (1997) ce sont plutôt des étiquettes caractérisant un état mental particulier qui fait agir l'agent de telle manière qu'on a l'impression qu'il ressent cette

émotion même si ce n'est pas vraiment le cas.

C.2.1 Pourquoi doter des agents virtuels d'émotions artificielles ?

La psychologie a reconnu le rôle des émotions dans le fonctionnement de l'esprit humain, ainsi que leur fonction adaptative. Certains psychologues ont même proposé des modèles facilement adaptables en informatiques. Pourtant les informaticiens ne s'intéressent d'abord aux émotions que pour rendre leurs agents plus crédibles. Cette notion de "believability" a été introduite par Bates (Section 2.2.1). Puis la neurologie a prouvé le rôle des émotions dans l'intelligence et créer des modèles de leur impact sur le raisonnement (Section 2.2.2). Plusieurs informaticiens prennent alors conscience de l'importance des émotions pas seulement pour la crédibilité de leur agent mais aussi à cause de leur impact sur les interactions avec l'utilisateur : c'est le début de l'"Affective Computing", introduit par Picard (Section 2.2.3). Les agents émotionnels deviennent alors à la mode et de nombreux travaux voient le jour à leur sujet. Dans la section 2.2 nous discutons les premiers travaux de différents domaines de recherche visant à intégrer les émotions dans des agents virtuels ou à prouver leur utilité pour ces agents.

C.2.2 Agents émotionnels et applications

Nous décrirons ensuite quelques architectures émotionnelles (Section 2.3), ainsi que leurs applications pour la création d'agents pédagogiques (Section 2.4) et d'agents conversationnels (Section 2.5).

Nous remarquons alors que dans ces applications, les émotions sont souvent directement codées dans les agents. Les modèles sont donc souvent *ad hoc*, spécifiques à leur application, non génériques et non réutilisables. C'est pourquoi on voit émerger un nouveau domaine de recherche, visant la création de modèles formels génériques des émotions. Dans la section 2.6 nous expliquons pourquoi les logiques BDI peuvent être utilisées pour créer de tels modèles et décrivons quelques essais dans ce sens.

C.2.3 Modèles formels des émotions en logique BDI

Nous avons différencié plusieurs types de travaux sur les émotions en informatique. D'abord certains chercheurs conçoivent des architectures émotionnelles qui simulent l'influence des émotions sur diverses fonctions cognitives ainsi que leur interaction avec l'humeur et la personnalité. L'agent EMA est doté de l'architecture la plus ocmplète, qui intègre même des stratégies de *coping*. Ces travaux correspondent à ce que Gratch et Marsella appellent les approches basées sur la simulation

"simulation-based approaches", *cf.* Section 2.2.4). Mais ces architectures sont souvent complexes et exigent des représentations spécifiques.

Ensuite d'autres chercheurs implémentent directement des émotions dans leurs agents pour atteindre un but spécifique comme les rendre plus crédibles durant une conversation avec l'utilisateur (Section 2.5) ou les faire motiver un étudiant (Section 2.4). Ces approches correspondent donc plutôt à ce que Gratch et Marsella appellent les approches guidées par la simulation ("simulation-driven approaches"). Ils sont peu intéressés par le fonctionnement des émotions et par leur influence sur les fonctions cognitives de l'agent (mémoire, coping...). Ils cherchent principalement à rendre leur agent plus crédible du point de vue de l'utilisateur, si bien qu'ils ne sont pas toujours très fidèles aux théories psychologiques.

De plus, même si certains agents sont basés sur une architecture émotionnelle existante (*e.g.* Steve et Herman qui intègrent l'architecture *Affective Reasoner*), la plupart ont leurs propres modules émotionnels, dépendants du domaines et donc ni génériques ni réutilisables. Les informaticiens se basent donc sur diverses théories psychologiques et utilisent diverses techniques pour formaliser ces théories. Ces efforts désorganisés rendent difficile la réutilisation de ces modèles.

Il semble donc que la communauté agent ait besoin d'une architecture émotionnelle générique et réutilisable, fidèle à la psychologie, et basée sur une technologie répandue prête à être implémentée dans des agents. Nous avons exploré l'utilisation des logiques BDI pour développer un tel modèle. Comme l'utilisation de la logique peut sembler surprenante voire contradictoire, nous commençons par exposer (in Section 2.6.1) la pensée de Searle selon laquelle les émotions sont des attitudes mentales ocmplexes. Nous discutons alors plusieurs tentatives de formaliser les émotions d'un agent dans un cadre BDI : le langage émotionnel de Meyer basé sur sa logique KARO (Section 2.6.2) et la théorie de l'interaction rationnelle de Sadek utilisée par Ochs *et al.* (Section 2.6.3). Comme notre but est aussi de procurer une formalisation logique des émotions, nous critiquerons les formalisations existantes et conclurons sur comment nous pensons les améliorer (Section 2.6.4).

C.2.4 Conclusion

Ce chapitre a prouvé que les émotions valent la peine d'être intégrées dans des agents virtuels, car elles améliorent leur comportement dans un grand nombre d'applications, qui vont des agents pédagogiques aux mondes virtuels. Nous avons alors remarqué l'existence d'un grand nombre de travaux désorganisés pour créer des agents émotionnels, ce qui rend nécessaire la création d'un modèle générique des émotions que les informaticiens pourront réutiliser dans divers agents pour diverses applications. Nous avons alors discuté diverses tentatives pour procurer un

tel modèle, en particulier en utilisant les logiques BDI, une technologie répandue pour décrire des architectures d'agents rationnels. La philosophie de l'esprit soutient l'hypothèse sous-jacente que les émotions peuvent être considérées comme un type d'attitudes mentales. Nous avons alors critiqué plusieurs défauts des formalisations BDI existantes des émotions, en particulier le manque de généricité dû à l'utilisation de buts au lieu de désirs, et l'accent mis sur les aspects individuels aux dépends des aspects sociaux.

Au contraire notre but dans cette thèse, comme nous l'avons dit en introduction, est de proposer un modèle formel des émotions. En accord avec Searle ou Meyer, nous considérons les émotions comme des attitudes mentales particulières et choisissons de baser notre modèle sur une logique BDI. Ce modèle devra remplir plusieurs contraintes. D'abord il devra être aussi fidèle que possible à la typologie OCC, la théorie psychologique que nous avons choisi de formaliser. En effet la psychologie a déjà expérimenté ses définitions depuis des décennies et nous devons donc leur faire confiance. Deuxièmement nous voulons formaliser un grand nombre d'émotions différentes, car des expressions émotionnelles diversifiées sont essentielles pour rendre un agent crédible. En fait nous proposerons des définitions pour vingt émotions de la typologie OCC. Troisièmement, nous voulons considérer les aspects sociaux des émotions et nous introduirons donc un opérateur modal pour représenter les normes sociales. Quatrièmement nous voulons rester aussi génériques que possible, en particulier notre modèle ne devra pas être orienté tâche comme c'est le cas pour ceux de Meyer et Ochs *et al.*: nous exprimerons donc les émotions en termes de désirs plutôt qu'en termes de buts (ce choix sera plus amplement discuté au Chapitre 4). Cinquièmement nous voulons être capables de raisonner au sujet des émotions et de prouver certaines propriétés, c'est pourquoi nous nous efforcerons de fournir une logique correcte et complète

La prochaine partie de cette thèse est dédiée au travail formel qui constitue le coeur de cette thèse : introduction du cadre logique (Chapitre 3), formalisation de la typologie OCC (Chapitre 4), et preuve formelle de propriétés des émotions (Chapitre 5).

C.3 Cadre logique

Comme nous l'avons montré dans l'état de l'art, la communauté agent a récemment commencé a s'intéresser de près aux agents émotionnelles. Malheureusement la plupart des approches conçoivent leur propre modèle émotionnel en partant directement des théories psychologiques. Il existe donc une grande variété de modèles informatiques des émotions, selon leur application, leur contexte d'utilisation, ou le formalisme sous-jacent (*cf.* Chapitre 2). Nous pensons que la communauté

agent a besoin d'un modèle formel générique des émotions, et nous soutenons que les logiques BDI (*viz*. logiques de la croyance, du désir et de l'intention, *e.g*. Cohen and Levesque (1990), Rao and Georgeff (1991; 1992), Sadek (1992), Herzig and Longin (2004)) permettent de développer un tel modèle.

En effet les émotions sont un phénomène complexe dont la définition psychologique est souvent abstraite, ambiguë et non consensuelle, ce qui mène à de nombreux débats (*cf*. Chapter 1). Les définitions psychologiques des émotions sont donc souvent subjectives alors que les informaticiens recherchent des définitions claires, prêtes à être implémentées dans leurs agents. La logique formelle procure un vocabulaire universel avec une sémantique claire. Elle permet aussi de raisonner, de planifier et d'expliquer le comportement d'un agent. La formalisation logique d'un phénomène peut même révéler des problèmes non évidents. Les logiques BDI (Cohen and Levesque (1990), Rao and Georgeff (1991,1992), Sadek (1992), Herzig and Longin (2004), Wooldridge (2000)), basées sur la philosophie du langage, de l'esprit et de l'action (*cf*. Bratman (1987), Searle (1969, 1983)), proposent de modéliser les agents grâce à quelques concepts clés comme l'action et les *attitudes mentales* (croyances, buts, désirs, intentions, choix...). Ce cadre est communément utilisé dans la communauté agent et offre des propriétés intéressantes bien connues : grand pouvoir explicatif, vérifiabilité formelle, cadre théorique rigoureux et bien établi d'un point de vue philosophique comme logique.

Dans ce chapitre nous fournissons donc un cadre logique BDI pour modéliser les émotions. Notre logique est une logique modale propositionnelle. Alors qu'il y a un large consensus sur la notion de croyance, il existe plusieurs concepts de désir dans la littérature. Ces concepts peuvent être rangés dans deux catégories. Dans la première le désir est vu comme quelque chose qui est abandonné une fois satisfait (comme le concept d'intention de Bratman), comme le désir de voir le soleil en un jour pluvieux, qui disparaît quand le soleil revient. Dans la seconde catégorie les désirs correspondent plutôt à des préférences générales dont l'existence ne dépend pas de leur satisfaction, comme la préférence générale pour les journées ensoleillées. Dans beaucoup d'approches BDI c'est cette seconde approche qui est choisie, et de plus les désirs sont fortement reliés aux croyances (*cf*. le concept de but de Cohen and Levesque (1990) ou Rao and Georgeff (1991), ou le concept de choix de Sadek (1992) ou Herzig and Longin (2004)). Ici nous choisissons aussi la deuxième vue, mais ne connectons pas les désirs aux croyances comme le font les autres approches. Finalement, il est apparu que le concept d'intention pouvait être omis dans notre cadre, ainsi que toutes les attitudes mentales "intermédiaires" (choix, but à atteindre, but persistant, intention...), car ils ne sont pas nécessaires pour décrire les émotions de la typologie OCC (cependant il faut noter que les in-

tentions seront essentielles pour décrire les stratégies de *coping*, (*cf.* Chapter 8)). De plus notre cadre utilise aussi des opérateurs modaux de temps, d'action, et d'idéalité. Notre notion d'idéalité permet de représenter les standards (moraux, légaux...) intériorisés par un agent.

Ce chapitre présente en détail notre cadre BDI, la sémantique et l'axiomatique de nos opérateurs modaux. De plus nous y prouvons la correction et la complétude de notre logique. Nous disposons alors d'un ensemble d'opérateurs modaux pour décrire les attitudes mentales et les capacités de raisonnement d'un agent : croyances, probabilités, désirs personnels, idéaux sociaux intériorisés, représentation du temps et de l'action. La prochaine étape est de caractériser les émotions en termes de ces concepts. La correction et la complétude de notre logique sera alors très importante pour raisonner formellement au sujet des émotions et prouver certaines de leurs propriétés.

C.4 Définitions formelles des émotions

C.4.1 Choix d'une théorie psychologique à formaliser

Dans ce chapitre nous nous attaquons au coeur du travail : formaliser les émotions. Nous avons déjà le cadre formel, il reste à justifier la source psychologique. Comme nous l'avons dit au Chapitre 1, les théories de l'évaluation cognitive sont mieux adaptées pour raisonner au sujet des émotions et de leurs antécédents cognitives, c'est-à-dire de l'état mental particulier qui les a causées. Parmi les diverses théories de l'*appraisal* disponibles, nous choisissons de formaliser la typologie OCC (1988, *cf.* Section 1.3.3). En effet elle est plus simple à comprendre pour des informaticiens que beaucoup d'autres théories car elle a été conçue pour être appliquée en Intelligence Artificielle. Cette approche est donc devenue la plus citée par la communauté agent. Bien sûr cet argument n'assure pas que cette théorie soit mieux adaptée pour développer des agents émotionnels. Ce problème sera discuté au Chapter 7 où nous présenterons une expérience d'évaluation des résultats donnés par cette théorie. En attendant dans ce chapitre nous partons de la typologie OCC et essayons de formaliser ses définitions dans notre langage logique présenté au Chapitre 3. En fait nous formalisons seulement deux branches de cette typologie. La branche des émotions déclenchées par des événements contient des émotions dont les conditions de déclenchement dépendent de l'évaluation de la désirabilité des conséquences d'un événement par rapport à leur impact sur les buts de l'agent. La branche des émotions déclenchées par les actions des agents contient des émotions dont les conditions de déclenchement dépendent du jugement de la responsabilité d'un agent (mérite ou blâme) qui a réalisé une certaine action, selon

son respect des normes sociales, ou plus exactement de celles d'entres elles qu'il a intériorisées, adoptées comme importantes.

C.4.2 Difficultés d'une formalisation logique des émotions

Une telle initiative soulève de nombreuses difficultés. La première est de comprendre ce que signifient les définitions psychologiques : celles-ci impliquent souvent des concepts ambigus, et nous devons les interpréter avant de les formaliser. La deuxième est de trouver une formalisation adaptée pour ces concepts : en effet un langage logique est inévitablement beaucoup moins expressif que le langage naturel, et il est difficile de représenter des concepts psychologiques complexes avec un jeu limité d'opérateurs modaux. Nos définitions formelles dépendent donc de notre interprétation des définitions informelles d'Ortony *et al*., et cette interprétation est bien sûr discutable. C'est pourquoi nous n'exposerons pas nos définitions formelles sans justifier soigneusement nos choix de formalisation. De plus nous soutiendrons la justesse de nos choix en montrant que nos définitions permettent de simuler les exemples cités par Ortony *et al*. pour illustrer leurs définitions.

C.4.3 Structure du chapitre

Dans ce chapitre nous exposons dix paires d'émotions opposées de la typologie OCC, en respectant toujours la même structure. Nous commençons par donner la définition psychologique ; puis nous proposons notre formalisation de cette définition et montrons qu'elle capture l'exemple donné par les auteurs. Nous discutons ensuite nos choix de formalisation et les appuyons encore par des exemples. Finalement nous comparons les définitions de ces émotions par Ortony *et al*. avec celles données par Lazarus ((Lazarus, 1991), Section 1.3.2) pour les mêmes émotions ou des émotions proches.

C.4.4 Limitations de notre formalisation

C.4.4.1 Intensité des émotions, dynamique, mélange

Tout d'abord nos opérateurs modaux ne sont pas valués : ils n'ont pas de degré associé, car il est très difficile d'associer une sémantique à de tels opérateurs (Laverny and Lang (2004,2005a)). Nos émotions ne sont donc pas valuées non plus, elles n'ont pas de degré d'intensité associé. C'est un défaut important car l'intensité est essentielle pour caractériser une émotion (Frijda et al., 1992). Cela nous empêche de formaliser des différenciations fines entre des émotions du même type (par exemple : irritation, colère, rage), pourtant cruciales pour un agent expressif. Cela nous empêche aussi de gérer leur évolution au cours du temps, en particulier

leur décroissance normale après leur apparition. Nos émotions persistent donc tant que leurs conditions restent vraies. Enfin nous ne pouvons pas non plus décrire le mélange de plusieurs émotions déclenchées simultanément ; Gershenson (1999) propose une solution originale à ce problème. De notre côté nous laissons ces problèmes de calcul d'intensité pour des travaux ultérieurs. Cependant quand nous avons implémenté notre agent (*cf*. Chapter 7) nous avons fait quelques concessions au niveau de la sémantique pour obtenir des émotions avec une intensité.

Dans leurs définitions Ortony *et al*. utilisent ce qu'ils appellent des *intensity variables* qui influencent l'intensité d'une émotion. Comme nous ne calculons pas l'intensité des émotions nous avons intégré ces variables dans la définition quand elles nous paraissaient très importantes, et nous les avons négligées sinon. Nous ne conservons ainsi que l'essence du type générique d'émotion, l'intervention des variables d'intensité créant des émotions particulières dans ce type.

C.4.4.2 Limitations à cause des opérateurs modaux

Un autre problème est la traduction des concepts complexes que peut exprimer le langage naturel dans un vocabulaire limité d'opérateurs modaux. Il peut paraître simple d'inventer des opérateurs qui correspondent exactement aux concepts nécessaires, mais ce n'est pas le cas. Comme nous voulions conserver une logique correcte et complète, nous avons dû faire quelques simplifications. C'est pourquoi nous utilisons une logique temporelle linéaire où manquent certains opérateurs qui auraient pu être utiles (*cf*. Section 4.4.3.1). Nous n'avons pas non plus pu représenter précisément le lien entre une action et son effet (*cf*. Remark 1), car notre logique $\mathbf{K}_t$ ne procure aucun opérateur qui exprime ce lien. L'opérateur $STIT$ fait actuellement l'objet de recherches (Chellas, 1992; Horty and Belnap, 1995b; Troquard, Trypuz, and Vieu, 2006) pour résoudre ce manque d'expressivité. Finalement ce travail exclut pour l'instant les émotions déclenchées par les aspects d'objets, car notre logique propositionnelle ne permet pas de les représenter. Dans des travaux futurs une logique modale des prédicats pourrait permettre de les décrire.

Malgré ces simplifications nous avons réussi à formaliser vingt émotions de la typologie OCC.

C.4.5 Comparaison entre Lazarus et OCC

Nous avons comparé dans ce chapitre les définitions offertes par Lazarus et par Ortony *et al*. de certaines émotions. Lazarus semble au premier abord offrir des définitions plus précuses, faisant intervenir des variables d'évaluations plus complexes pour différencier de manière fine des émotions que Ortony *et al*. regroupe

dans le même type d'émotion (*e.g*. culpabilité et honte, envie et jalousie, angoisse et anxiété). Mais d'un autre côté il néglige certaines émotions définies par Ortony *et al*. (*e.g*. admiration, reproche, remords, content pour, *gloating*...) et qui semblent importantes. Finalement nous pouvons dire que la typologie OCC paraît plus adaptée pour concevoir un agent émotionnel expressif. En effet on attend d'un tel agent qu'il exprime des émotions adaptées dans une grande variété de situations, il doit être robuste. Au contraire Lazarus privilégie la précision par rapport à la robustesse, si bien qu'un agent basé sur cette théorie pourrait exprimer des émotions très précises dans certaines situations, mais n'exprimerait aucune émotion dans d'autres. La théorie de Lazarus nous semble donc pouvoir être utile seulement dans un deuxième temps : si les émotions de l'agent ne sont pas assez précises dans certains cas nous pourrons raffiner nos définitions en nous inspirant de cette théorie. Nous vérifierons par la suite si nous avons besoin de telles améliorations en évaluant la crédibilité de notre agent auprès d'utilisateurs humains (*cf*. Chapter 7).

C.4.6 Suite du travail

Une fois que nos définitions formelles sont acceptées, nous pouvons prouver certaines propriétés des émotions comme théorèmes de notre logique. Dans le Chapitre 5 nous exposons et prouvons quelques propriétés, en particulier des liens causaux et temporels entre certaines émotions.

Ces propriétés contribuent à soutenir la justesse de nos définitions formelles. Pour évaluer plus précisément notre modèle nous avons décidé de l'implémenter dans un agent BDI. Cet agent expressif a ensuite été soumis à l'évaluation d'humains qui ont jugé la pertinence et la crédibilité de ses émotions pendant un court scénario (*cf*. Chapter 7). Ce travail a donné des résultats encourageants et nous projetons de renouveler ces expériences en coopération avec des psychologues qui pourraient tirer parti de ce moyen pour évaluer leurs théories.

Finalement nous pouvons remarquer que ce travail de formalisation ne concerne pour l'instant que le processus d'*appraisal*, qui conduit au déclenchement des émotions. Mais pour les psychologues le mécanisme émotionnel comprend aussi un processus de *coping* qui décrit l'influence des émotions sur le comportement. Nous travaillons actuellement sur la formalisation de ce processus dans le même cadre BDI. Le chapitre 8 décrit l'état actuel de notre recherche à ce sujet.

C.4.7 Conclusion

Un tel travail offre donc de nombreuses perspectives de continuation qui peuvent être intéressantes pour différentes branches de recherche. Les concepteurs d'agents

conversationnels animés seront intéressés par donner des capacités de *coping* à leur agent pour rendre son comportement émotionnel et crédible (*cf*. Chapter 8). Les psychologues trouveront intéressant d'évaluer leurs théories grâce à un agent émotionnel (*cf*. Chapter 7). Enfin les philosophes et les logiciens apprécieront la possibilité de prouver formellement des propriétés des émotions. C'est l'objet du prochain chapitre.

C.5 Propriétés formelles des émotions

C.5.1 Introduction

Comme nous l'avons vu en décrivant les théories psychologiques des émotions (Chapitre 1), les émotions sont un phénomène complexe qui a toujours été l'objet de débats. En effet la définition et les propriétés de concepts définis de manière informelle sont toujours discutables. Au contraire une formalisation logique d'un concept permet de le désambiguïser.

Dans ce chapitre nous montrons comment notre formalisation logique des émotions (Chapitre 4) permet de les désambiguïser et de raisonner à propos de leurs propriétés, à condition que nos définitions soient acceptées. En effet comme notre logique est correcte et complète, si certaines propriétés des émotions sont vraies nous devons pouvoir les prouver, et si nous pouvons prouver certaines propriétés c'est qu'elles sont vraies. Ces propriétés ne seront alors plus discutables.

Ce chapitre expose donc et prouve quelques propriétés des émotions, en particulier concernant les relations temporelles et causales qu'elles entretiennent (ou n'entretiennent pas) les unes avec les autres. Certaines de ces propriétés vont au-delà de ce qu'Ortony *et al*. exposent dans leur livre mais elles demeurent intuitives. Un tel travail souligne les atouts d'un raisonnement formel sur les émotions.

C.5.2 Conclusion

Ce travail a contribué à montrer l'intérêt des logiques BDI pour formaliser les émotions. D'abord, il existe déjà un grand nombre de travaux sur ces logiques dans la communauté agent, et un tel modèle est prêt à être utilisé dans de nombreux agents existants. De plus nous avons montré qu'une fois nos définitions acceptées nous pouvions prouver un grand nombre de propriétés intuitives des émotions. Seule une formalisation logique peut offrir de tels résultats sans équivoque au sujet de phénomènes qui ne sont pas toujours clairement analysés dans la littérature psychologique.

Cependant, les logiques BDI ne sont pas seulement un outil permettant de faire des démonstrations. Elles sont aussi un puissant moyen de décrire le raisonnement

d'un agent. Dans le prochain chapitre nous détaillons deux applications où notre modèle intégré dans un agent BDI permet d'améliorer son fonctionnement.

C.6 Applications

C.6.1 Introduction

Les agents émotionnels trouvent de plus en plus d'applications (*cf.* Chapter 2) dans des domaines aussi variés que les jeux vidéos, les logiciels pédagogiques, l'apprentissage assisté par ordinateur, les environnements virtuels d'entraînement, les agents conversationnels animés, les compagnons virtuels... Dans ce chapitre nous décrivons deux applications que nous avons explorées pour notre modèle des émotions. Ces applications ont été réalisées à un moment où notre modèle n'était pas encore aussi riche que dans sa version actuelle, et ne couvrent donc pas toutes les émotions définies au chapitre 4. Elles illustrent néanmoins qu'un tel modèle n'est pas seulement utile pour raisonner abstraitement sur les émotions mais aussi pour développer des agents émotionnels pour des applications concrètes. Ce chapitre en décrit deux parmi les nombreuses possibles : un agent intelligent conscient des émotions de l'utilisateur afin de prendre soin de lui dans le cadre d'une application d'Intelligence Ambiante ; et un agent conversationnel qui exprime ses émotions pendant un dialogue avec un autre agent, pour paraître plus crédible et favoriser l'immersion de l'utilisateur dans un monde virtuel.

C.6.2 Intelligence Ambiante

L'Intelligence Ambiante est l'art de concevoir des environnements intelligents, *viz.* des environnements qui peuvent à tout moment adapter leur comportement à leur utilisateur, à ses buts et ses besoins spécifiques, de manière à assurer son bien-être de manière non intrusive voire quasiment invisible. Notre première application consiste à concevoir un agent BDI capable de gérer ces tâches, et donc conscient des émotions de l'utilisateur.

Les concepteurs d'agents ont exploré diverses méthodes pour découvrir l'émotion de l'utilisateur quand celui-ci ne l'exprime pas directement. Prendinger and Ishizuka (2005) utilisent les résultats de Picard (1997) pour déduire l'émotion de l'utilisateur à partir du suivi de ses signaux physiologiques et de la direction de son regard. Cette méthode permet de détecter en temps réel le moindre changement dans l'émotion du sujet mais elle est plutôt envahissante, ce qui est contraire à un principe important de l'Intelligence Ambiante. Une autre méthode a été explorée par Jaques et al. (2004). Leur agent pédagogique déduit l'émotion de son élève en se mettant à sa place (grâce à un modèle de l'utilisateur) pour évaluer les événements via une

fonction d'*appraisal* basée sur la typologie OCC. Cette méthode ne fournit qu'une vue subjective de l'état émotionnel de l'utilisateur mais elle est plutôt efficace si elle est couplée à un bon modèle de ses attitudes mentales. De plus elle n'est pas du tout envahissante, ce qui est très important. Par ailleurs cette méthode permet aussi de raisonner sur les émotions, de comprendre par exemple leurs causes, et est donc plus adaptée à la problématique de l'Intelligence Ambiante.

Une fois que l'agent connaît l'émotion de l'utilisateur il peut aider celui-ci à y faire face si elle est négative (*cf.* Chapter 8). En psychologie (Lazarus and Folkman, 1984) le *coping* est le processus de choix par l'agent d'une stratégie visant à supprimer ou atténuer une émotion négative qu'il ressent (par exemple en minimisant ou supprimant ses causes). Nous considérons ici qu'un agent émotionnel pour l'Intelligence Ambiante peut aider l'utilisateur dans cette tâche.

Finalement nous pensons que pour un agent intégré dans un Système d'Intelligence Ambiante (SIA), un modèle informatique des émotions est utiles dans les cas suivants :

- *(C1)* pour calculer l'émotion déclenchée chez l'utilisateur par un événement extérieur (comme dans (Jaques et al., 2004)) ;
- *(C2)* pour anticiper l'effet émotionnel des actions possibles de l'agent sur l'utilisateur :
  - soit pour choisir une action dans le but de produire un certain effet émotionnel *(C2a)*
  - soit pour choisir entre plusieurs actions ayant des effets physiques comparables *(C2b)* ;
- *(C3)* pour comprendre les causes d'une émotion que le comportement de l'utilisateur semble exprimer. Cette explication peut se faire :
  - soit de manière directe quand toutes les informations nécessaires sont connues *(C3a)*
  - soit via la formulation d'hypothèses sur les croyances de l'utilisateur *(C3b)*.

  Connaître l'émotion ressentie par l'utilisateur ainsi que les causes de cette émotion est fondamental pour agir de manière réellement adaptée.

Dans cette première application nous utilisons notre modèle des émotions pour décrire le comportement d'un agent intégré dans une SIA contrôlant une maison intelligente chargé de prendre soin de son habitant en l'aidant à gérer ses émotions. Nous montrons la puissance de notre cadre logique sur cinq scénarios différents

correspondants aux cinq cas d'utilisation identifiés ci-dessus. Dans chaque cas nous considérons que la maison intelligente est administrée par l'agent m, qui peut recevoir de l'aide d'autres agents du SIA. Nous appelons h l'habitant de cette maison intelligente.

Un tel raisonnement peut être utile en Intelligence Ambiante, mais aussi pour les Interfaces Homme-Machine ou les agents pédagogiques. Seule une théorie cognitive des émotions permet de raisonner de cette manière sur les antécédents des émotions. Les logiques BDI sont alors particulièrement adaptées pour permettre à l'agent de réaliser ce raisonnement.

C.6.3 Dialogue

Notre modèle logique des émotions décrit leur déclenchement cognitif, *viz.* il définit les attitudes mentales qui les causent. Dans cette application nous nous intéressons plus particulièrement au déclenchement des émotions au cours du dialogue, i.e. nous considérons les émotions déclenchées par un énoncé, à la fois chez le locuteur et chez l'auditeur. Cependant nous ne décrivons pas l'expression de ces émotions (par des actes de discours expressifs par exemple) ni leur effet sur la suite du dialogue (par exemple une modification de la construction du dialogue).

Un dialogue peut être vu comme une séquence d'actes de discours (Austin, 1962; Searle, 1969; Searle and Vanderveken, 1985) réalisés par chaque énoncé. Ces actes de discours sont des actions particulières et on peut donc considérer qu'elles déclencheront les émotions basées sur les actions de la typologie OCC. Elles seraient alors évaluées par rapport aux normes conversationnelles auxquelles les interlocuteurs sont soumis (par exemple les maximes de Grice (Grice, 1957)). Mais cela n'est pas suffisant pour évaluer un acte de discours de manière complète, car il transmet aussi des informations sur le monde qui doivent être évaluées. Par exemple, imaginons un enfant qui casse un vase précieux mais l'avoue spontanément à son père. Si le père n'évalue que l'action d'avouer une faute spontanément, il sera fier de son fils pour son courage et son honnêteté. Mais il évaluera certainement aussi l'information "ton vase précieux est cassé" (ce que nous appelons ici le *contenu propositionnel*) et sera triste de cette mauvaise nouvelle.

Recevoir un acte de discours est donc un autre moyen d'observer son environnement et de connaître les stimuli pertinents. Nous considérons la réception d'un acte de discours comme une perception indirecte d'un stimulus (celui qui est décrit dans son contenu propositionnel, même si l'acte de discours n'est pas un assertif), et ce stimulus peut déclencher une émotion de n'importe laquelle des trois branches de la typologie OCC. En fait nous nous restreignons dans cette application précoce aux cas où l'acte de discours est une assertion ou une requête au sujet d'un événement, et nous ne décrivons pas l'évaluation de l'action sous-jacente par rapport aux

normes conversationnelles.

Dans ce chapitre nous analyserons un exemple de dialogue émotionnels entre deux avatars[1]. Le contexte est un monde virtuel pour l'entraînement des pompiers.

Cette application montre que notre formalisation BDI des émotions peut être combinée avec une sémantique BDI des actes de langage pour décrire les émotions déclenchées pendant un dialogue. Cette application précoce est encore incomplète et a nécessité des simplifications et des hypothèses, mais elle pourra facilement être étendue à d'autres cas. Les émotions en réaction à la description d'une action sont déclenchées de manière similaire que pour la description d'un événement. Les émotions en réactions à des actes de discours considérés comme des actions dépendent de normes conversationnelles. Une fois celles-ci formalisées, ces émotions découleront naturellement de nos définitions. La formalisation de ces normes constitue une continuation intéressante de ce travail.

De plus cette description est limitée à l'expression des émotions durant le dialogue, ces émotions n'ayant encore aucun effet sur le dialogue. Nous pensons que cette influence pourra être exprimée en termes de stratégies de *coping*, *viz*. les efforts qu'un individu fait pour gérer ses émotions. Une continuation naturelle de ce travail est donc de formaliser ces stratégies de *coping* dans notre cadre BDI ((*cf*. Chapter 8)).

C.6.4 Conclusion

Nous avons déjà discuté les avantages d'un modèle logique des émotions. Ce chapitre montre qu'un tel modèle n'est pas seulement utile pour réaliser des preuves formelles des propriétés des émotions, mais qu'il est aussi fonctionnel et prêt à être implémenté et utilisé pour divers buts. Nous n'en avons exploré que deux mais nous pensons qu'il en existe beaucoup plus (*cf*. Chapter 2).

Par ailleurs ces deux applications révèlent l'incomplétude de notre modèle actuel des émotions. Nous avons formalisé le déclenchement des émotions mais pas leur influence sur le comportement. Ainsi les exemples précédents n'ont pu être que partiellement formalisés. Notre première tentative pour corriger ce défaut en formalisant les stratégies de *coping* sera exposé au chapitre 8.

[1] Nous avons implémenté une plate-forme permettant la génération automatique de tels dialogues émotionnels entre deux agents conversationnels. Cette plate-forme nommée a été présentée lors d'un workshop francophone (Adam and Evrard, 2005)

C.7 Implémentation et évaluation

Comme nous l'avons montré dans l'introduction de ce mémoire, la communauté agent conçoit de plus en plus d'agents émotionels, en particulier des agents conversationnels animés (*cf.* Chapter 2). Pour garantir la pertinence des émotions exprimées dans le contexte de l'interaction (et ainsi préserver la crédibilité de l'agent) les chercheurs s'appuient sur des théories psychologiques (*cf.* Chapter 1), et plus particulièrement sur la typologie OCC (Ortony, Clore, and Collins, 1988)). Nous pensons que la principale raison de cet engouement est la simplicité de cette typologie, qui la rend très accessible pour des informaticiens. Par contre nous remarquons que d'un point de vue psychologique rien ne prouve que cette théorie soit meilleur que les autres. Dans le cadre de la modélisation d'agents aussi crédibles et réalistes que possible cette question est pourtant essentielle, notamment du fait que les théories psychologiques divergent parfois sensiblement dans leurs définitions des émotions (*cf.* Chapter 1).

Dans ce chapitre nous voulons donc évaluer la pertinence des émotions qu'un agent BDI peut exprimer en utilisant notre formalisation de la typologie OCC. Pour cela nous implémentons notre cadre BDI dans un agent virtuel nommé PLEIAD (*Prolog Emotional Intelligent Agents Designer*)[2]. PLEIAD exprime les émotions qu'il "ressent" en réponse aux stimuli envoyés par l'utilisateur. Nous tenons à préciser dès maintenant que PLEIAD intègre à titre expérimental la gestion de degrés numériques associés aux attitudes mentales, and de degrés d'intensité numériques associés aux émotions. Nous avons alors testé PLEIAD sur un court scénario et demandé à quelques personnes d'évaluer la pertinence des émotions exprimées par l'agent (correspondent-elles aux émotions qu'ils auraient eux-mêmes ressenties dans la même situation ?) et leur crédibilité (correspondent-elles à des émotions communément admissibles dans ce genre de situations ?). Notre but est:

- de tester les prédictions de notre modèle théorique en les comparant aux attentes de l'utilisateur ;
- de tester les prédictions de la typologie OCC elle-même par rapport à ces attentes.

Nous décrivons d'abord l'implémentation de l'agent PLEIAD (Section 7.2), puis exposons les modalités de de notre évaluation (Section 7.3), et finalement discutons les résultats des évaluations et donnons nos conclusions au sujet de notre modèle mais aussi au sujet de la théorie psychologique sous-jacente (Section 7.4).

[2]PLEIAD a été présenté pour la première fois lors du Workshop Francophone WACA'2006, *cf.* (Adam, 2006). Les résultats de cette évaluation seront publiés dans une revue française (Adam, Herzig, and Longin, 2007).

Les résultats de cette première évaluation ouvrent de nombreuses perspectives d'amélioration, au moins pour la partie qui dépend de nous, *viz.* notre propre formalisation de la typologie OCC. Notre agent pourrait permettre de comparer plusieurs formalisations différentes de cette typologie, à condition qu'elles soient toutes exprimées dans la même logique. Par ailleurs il permet aussi de comparer les prédictions de plusieurs théories psychologiques des émotions, à condition aussi de les formaliser dans la même logique. Cette formalisation représente cependant un gros travail qui pourra faire l'objet de recherches futures.

Nous nous intéressons en particulier à la théorie de l'*appraisal* de Lazarus, qui a souvent semblé plus subtile et plus exacte que la typologie OCC durant les évaluations. Cependant cette théorie est beaucoup plus complexe que la typologie OCC, en partie car elle n'a pas été conçue pour être utilisée en Intelligence Artificielle. Elle implique des concepts complexes de responsabilité, d'implication de soi ("ego-involvement")... qui seront difficile à formaliser dans notre logique, et plus généralement dans tout formalisme qui ne soit pas trop complexe. Pour formaliser cette théorie, Gratch and Marsella (2004a) ont créé leur propre structure complexe de représentation de l'état mental de l'agent, en adaptant divers formalismes existants (*cf.* Chapter 2). Un moyen plus simple est de représenter la responsabilité via le concept d'"agency" que nous pouvons intégrer dans notre logique BDI grâce à l'opérateur STIT (*seeing-to-it-that*, (Horty and Belnap, 1995a)), que nous avons déjà commencé à étudier (*cf.* (Herzig and Troquard, 2006; Broersen, Herzig, and Troquard, 2006)).

Cependant nous devons nous poser la question du rapport entre le profit réalisé en termes d'expressivité et crédibilité et les coûts additionnels produits par la complexification de notre théorie. Est-il vraiment nécessaire pour un agent d'exprimer des différences subtiles entre la culpabilité et la honte, ou entre la jalousie et l'envie ? Cela dépend grandement de l'application pour laquelle l'agent est prévu (*cf.* Chapter 2). Finalement la théorie émotionnelle idéale pour ces agents pourrait être un compromis entre plusieurs théories, utilisant parfois des notions simples mais suffisantes et parfois des notions plus complexes pour certaines émotions critiques. Mais de toutes façons, du fait que les chercheurs tentent de rendre leurs agents aussi crédibles que possible, nous pensons qu'ils ne peuvent pas se permettre d'ignorer des théories psychologiques plus complexes que la traditionnelle typologie OCC. De plus la psychologie elle-même pourrait tirer avantage de telles recherches et en particulier de la possibilité d'évaluer leurs théories, de manière à mieux comprendre les émotions humaines.

C.8 Vers une formalisation du processus de coping

C.8.1 Introduction

Les émotions prennent une place de plus en plus importante dans la conception d'agents pour diverses applications (*cf.* Chapter 2) : Intelligence Ambiante, agents crédibles pour les mondes virtuels ou les jeux vidéos, agents conscients des émotions de l'utilisateur pour des applications pédagogiques, ludiques, d'assistance ou d'interface.

Les travaux existants sur les émotions en informatique (y compris les nôtres) s'intéressent principalement au déclenchement des émotions (Pelachaud et al., 2002; Meyer, 2004). La plupart du temps les chercheurs procurent une implémentation de la typologie OCC. Pourtant cette typologie ne décrit que l'*appraisal* (le processus évaluant l'environnement de l'agent pour déclencher une émotion appropriée), un processus qui ne représente qu'une partie du mécanisme émotionnel humain. En effet, Lazarus a montré qu'un deuxième processus intervient : le *coping*. Au sens de Lazarus and Folkman (1984), le *coping* représente les tentatives conscientes de l'individu pour gérer des stimuli menaçants signalés par des émotions négatives intenses déclenchées par le processus d'*appraisal*. Très peu de modèles de ce second processus existent (Dastani and Meyer, 2006; Gratch and Marsella, 2004a; Marsella and Gratch, 2003), alors que de nombreuses études psychologiques avalisent l'influence des émotions sur le comportement (Damasio, 1994; Forgas, 1995).

Pourtant des agents avec des capacités de *coping* pourraient trouver des applications dans de nombreux domaines (*cf.* Chapter 6). Par exemple, les agents conversationnels animés pourraient non seulement exprimer leur émotion par une expression faciale ou des modifications vocales, mais la manifester par une modification complète de leur comportement dialogique. En effet nous pensons que certains comportements dialogiques humains sont en fait la manifestation de l'utilisation stratégies de *coping*: mentir, interrompre son interlocuteur, refuser de répondre... Les modèles standards du dialogue sont incapables d'expliquer de tels comportements qui bien que réalistes sont souvent considérés "irrationnels", car ils adoptent des hypothèses de rationalité (trop) fortes et restrictives qui ne correspondent pas au raisonnement humain. Un modèle des stratégies de *coping* pourrait corriger ce problème. Un autre exemple est l'Intelligence Ambiante : un agent intelligent prenant soin de l'utilisateur l'assiste en fait dans l'utilisation de stratégies de *coping*.

Dans ce chapitre nous proposons donc de formaliser quelques stratégies de *coping*. Ce travail est encore en cours et les résultats présentés ici sont préliminaires. Notre but n'est pas de fournir un modèle accompli du processus de *coping*

mais plutôt de désambiguïser les concepts impliqués dans l'implémentation d'un agent émotionnel afin de pouvoir raisonner à leur sujet. Nous avons déjà montré les avantages des logiques BDI pour désambiguïser des concepts complexes (*cf.* Chapter 4) et pour raisonner sur leurs propriétés (*cf.* Chapter 5). Étant donnée la complexité du processus émotionnel et l'expressivité limitée des logiques BDI un tel modèle est évidemment réducteur, mais il offre des avantages indéniables. Nous nous basons donc sur notre cadre BDI conçu pour formaliser l'*appraisal* (*cf.* Chapter 3) et nous l'étendons pour rendre compte du *coping*. En particulier nous avons besoin d'opérateurs de choix et d'intention qui n'étaient pas présents dans ce formalisme.

Nous adaptons alors le modèle COPE (Carver, Scheier, and Weintraub, 1989) qui propose un ensemble de quinze stratégies de *coping*. Nous considérons ces stratégies comme des actions dont les préconditions et les effets sont exprimés en termes des attitudes mentales de l'agent (croyances, désirs et intentions). Afin de simplifier cette première approche nous nous restreignons dans un premier temps au stratégies de *coping* concernant des émotions déclenchées par des événements. Les émotions affectent alors le comportement de l'agent de deux manières : directement par le choix et l'application d'une stratégie de *coping*, et indirectement par la modification des attitudes mentales impliquées dans son raisonnement.

Ce chapitre commence par une introduction du concept psychologique de *coping* (Section 8.2). Il décrit brièvement la sémantique et l'axiomatique des nouveaux opérateurs modaux que nous ajoutons à notre cadre logique (Section 8.3), tout en maintenant sa correction et sa complétude. Il définit ensuite de manière formelle quelques stratégies de *coping* dans ce cadre (Section 8.4), et illustre leur utilisation effective sur un exemple provenant d'une simulation d'entraînement pour les pompiers (Section 8.5). Finalement il présente une discussion d'autres formalisations existantes de stratégies de *coping*: l'agent EMA de Gratch and Marsella, l'Affective Reasoner d'Elliott, et le langage agent de Meyer (Section 8.6).

C.8.2 Conclusion

Cette thèse a finalement permis de montrer que les logiques BDI sont assez expressives pour décrire le processus émotionnel dans son intégralité, c'est-à-dire à la fois l'*appraisal* (*cf.* Chapter 4) et le *coping* (dans ce chapitre).

Les logiques BDI sont un cadre communément utilisé qui offre des propriétés intéressantes, principalement sa réutilisabilité dans un grand nombre d'agents existants basés sur une architecture BDI. Les domaines d'application pour de tels agents concernent principalement la conception de personnages humanoïdes pour les mondes virtuels : un modèle émotionnel plausible améliore leur crédibilité et donc l'immersion de l'utilisateur dans le monde virtuel. La possibilité de raisonner

à propos des émotions d'un autre agent ouvre aussi des applications en Intelligence Ambiante, où de tels agents pourraient détecter l'émotion de l'utilisateur pour l'aider à y faire face en lui proposant des stratégies de *coping* adaptées (*cf.* Chapter 6).

Il faut cependant mentionner quelques défauts de notre modèle. Tout d'abord nous ne formalisons pas toutes les stratégies de *coping* mais seulement celles que nous avons jugées les plus intéressantes pour des agents intelligents, et seulement concernant les émotions déclenchées par des événements. Deuxièmement nous ne gérons pas l'intensité et la dynamique des émotions, un problème trop important pour être résolu ici, et nous supposons donc que l'exécution d'une stratégie de *coping* fait simplement disparaître l'émotion immédiatement, au lieu de faire diminuer son intensité. Troisièmement, nous avons seulement esquissé les déductions formelles dans ce chapitre, et n'avons pas complètement élaboré les mécanismes de révision et de préservation des croyances en jeu. En particulier, le *coping* nécessite d'abandonner nos axiomes de préservation des désirs.

Nos perspectives à court terme après cette description préliminaire consistent à étendre ce modèle pour expliquer les émotions de la branche agent de la typologie OCC. À plus long terme nous envisageons d'implémenter ce modèle dans un agent conversationnel animé, afin de simuler des comportements dialogiques souvent observés dans les interactions humaines mais pas encore capturés par les modèles de dialogue actuels (par exemple pourquoi changeons-nous brusquement de sujet ou refusons-nous de répondre à une question ou d'admettre l'évidence). En effet nous pensons que de tels comportements révèlent l'utilisation de stratégies de *coping*.

Conclusion

Cette thèse a présenté un projet pluridisciplinaire, qui a commencé par la compréhension de la définition psychologique des émotions, puis proposé leur formalisation logique et la déduction de certaines propriétés, pour arriver finalement à l'implémentation informatique d'un agent émotionnel. Ce travail apporte de multiples contributions.

D'abord il offre à la communauté agent un modèle formel d'un grand nombre d'émotions. Nos définitions des émotions sont voulues le plus réalistes possibles, grâce à différents moyens : elles sont fidèles aux définitions psychologiques et peuvent rendre capturer les situations déclenchantes décrites dans la théorie originale ; elles permettent de prouver des propriétés intuitives des émotions ; et finalement des humaines ont évalué positivement la pertinence des émotions exprimées par un agent utilisant ces définitions. Ce modèle formel des émotions permet donc aux chercheurs d'intégrer des émotions dans leurs agents sans devoir eux-mêmes inter-

préter et formaliser une théorie psychologique. En effet cette tâche est difficile et ne devrait être faite qu'une fois avant d'être ensuite réutilisée. C'est pourquoi nous avons aussi fait en sorte que notre formalisation soit aussi générique que possible, en utilisant un cadre bien connu : les logiques BDI.

Ensuite ce travail souligne que les logiques BDI sont un outil puissant pour désambiguïser des concepts complexes et raisonner au sujet de leurs propriétés. En effet nous avons été capables de donner des définitions formelles et donc non ambiguës de vingt émotions, mais aussi de prouver des théorèmes au sujet de leurs liens entre elles. Les logiques BDI sont souvent critiquées car on croit leur utilité limitée à la réalisation de preuves certes formelles, mais trop abstraites. Pourtant ces logiques permettent aussi de décrire l'architecture d'un agent. Nous avons exploré deux applications pour un agent doté de notre modèle des émotions (un agent conversationnel et un système d'Intelligence Ambiante) et avons même implémenté un tel agent.

Cependant nos applications demeurent incomplètes, du fait que notre formalisme lui-même est incomplet. En effet le processus émotionnel humain est constitué de deux composants : l'*appraisal*, qui mène au déclenchement des émotions, et le *coping*, qui conduit à l'adaptation du comportement à l'émotion ressentie. Dans cette thèse nous avons uniquement formalisé le processus d'*appraisal*, si bien que les émotions déclenchées n'ont aucune influence sur le comportement de l'agent. C'est pourquoi le dernier chapitre procure un aperçu de nos travaux en cours sur le *coping*. Nous avons essayé de formaliser ce processus dans le même cadre BDI. Ce processus a encore fait l'objet de moins de travaux que le déclenchement ou l'expression des émotions. Cette thèse ouvre donc d'intéressantes perspectives de recherche future.

Notre travail peut de plus être encore amélioré sur plusieurs points. Une première amélioration locale concerne notre couverture de la typologie OCC, qui n'est pas complète : nous n'avons pas formalisé la branche des émotions déclenchées par les aspects d'objets, car notre logique actuelle ne permet pas de décrire ces aspects. Plus généralement, l'expressivité limitée des logiques BDI (par comparaison avec le langage naturel) implique plusieurs limitations de notre modèle. Tout d'abord nous ne pouvons pas calculer l'intensité des émotions déclenchées, ce qui nous empêche aussi de gérer leur décroissance au cours du temps ou leur mélange entre elles. Deuxièmement nous n'avons pas décrit l'interaction des différents phénomènes émotionnels entre eux, c'est-à-dire comment plusieurs émotions peuvent se combiner pour former l'humeur, et comment l'humeur ou les émotions courantes peuvent biaiser le déclenchement d'une nouvelle émotion. En effet ces interactions sont liées avec l'intensité des émotions impliquées. Troisièmement nous avons seulement pu donner une approximation du lien entre une action et ses effets, car aucun opérateur modal rendant compte de ce lien n'est encore com-

plètement axiomatisé. De plus nous n'avons pas exploré le concept d'émotions de groupe, qui selon nous ferait intervenir des attitudes mentales de groupe et nécessiterait donc l'introduction de nouveaux opérateurs non standards. Cependant nous avons commencé à formaliser les croyances de groupe (Tuomela, 1992; Gaudou, Herzig, and Longin, 2007) et prévoyons de formaliser aussi les idéaux de groupe. Enfin, l'étude de la complexité de notre logique dépasse le cadre de ce travail.

Plus généralement, nous avons restreint notre travail aux aspects cognitifs des émotions, et avons donc négligé les aspects biologiques, physiologiques et socio-culturels pourtant essentiels. Ainsi la culture et les normes sociales exercent une influence reconnue sur l'expression ou l'inhibition des émotions déclenchées. De plus nous avons seulement effleuré le problème du *coping* qu'il est pourtant crucial de formalisé pour rendre compte des émotions de manière exhaustive.

Cependant, malgré ces limitations, notre évaluation montre que les émotions que notre modèle peut simuler sont perçues comme plutôt crédibles. Nous supposons maintenant que la crédibilité est un aspect d'une propriété plus générale des systèmes d'interaction : ils doivent inspirer confiance. Castelfranchi, Falcone, and Marzo (2006) ont montré que les utilisateurs doivent avoir confiance en un système pour l'utiliser. Nous sommes actuellement impliqués dans un projet dont le but est de formaliser la notion de confiance dans une logique BDI. Cela devrait permettre de désambiguïser cette notion et de raisonner automatiquement à son sujet. Ce modèle sera alors implémenté dans un agent et testé. Comme cette thèse, il s'agit encore d'un projet pluridisciplinaire qui démarre de l'analyse philosophique et sociologique d'un concept et mène à sa formalisation et son utilisation dans un système informatique.

Finalement ce travail est juste un premier pas sur le long chemin qui mène à la compréhension des émotions. Ce chemin doit être exploré collectivement par des chercheurs provenant de disciplines aussi variées que la psychologie, l'informatique, la biologie, la sociologie... Cette coopération peut paraître difficile au premier abord, mais c'est ce qu'il en coûte de comprendre ce qui nous rend humains.