Chapter 1
The Cognitive Anatomy and Functions of Expectations Revisited

Emiliano Lorini

Abstract Some years ago (in 2003) I wrote my first paper in collaboration with Cristiano Castelfranchi who at that time was the supervisor of my Master project in Cognitive Science. The work was aimed at providing a logical formalization of the notion of expectation and of expectation-based emotions such as hope, fear, disappointment and relief. In this paper I will revisit and extend the analysis we did in 2003. I will propose a refinement of the notion of expectation by formalizing its two primitive components: the value of the goal and the strength of the belief. Thanks to this refinement, I will provide a formal analysis of the intensity of hope and fear.

1.1 Introduction

Cristiano Castelfranchi has been a great teacher and a constant source of inspiration. During the years of my PhD in Rome he taught me to love scientific research and how to find beautiful connections among different aspects of human mind and of social interaction. We had very intense discussions on several different topics: on the relationships between beliefs and motivations (*e.g.* goals, intentions *etc.*); on the concepts of intentional action and attempt; on the theory of altruism and pro-social attitudes; on the theory of social trust, power, delegation; on the cognitive theory of surprise, and on many many others. In 2003 I wrote my first scientific paper together with him who at that time was the supervisor of my Master project in Cognitive Science. The work - published in [7][1] - was aimed at providing a logical formalization of the notion of expectation and of expectation-based emotions such as hope, fear, disappointment and relief.

In this paper I will revisit and extend the analysis we did in 2003 into two directions. First of all, I will introduce a new logical framework which

Université de Toulouse, IRIT-CNRS, France

[1] A longer version of this work can be found in [8].

allows to formalize in a simple way some basic concepts for a formal theory of expectation-based emotions: the notion of graded belief and the notion of graded goal. Secondly, I will propose a refinement of the notion of expectation by formalizing its two primitive components: the value of the goal and the strength of the belief. Thanks to this refinement, I will provide a formal analysis of the cognitive structures of hope and fear and of their intensities. With cognitive structure of an emotion, I mean the emotion's triggering conditions, that is, the agent's mental states (beliefs, goals, intentions, *etc.*) that trigger the agent's emotional reaction (*e.g.*, action tendencies and physiological reactions) and 'cause' the agent to feel the emotion.

The rest of the paper is organized as follows. Section 1.2 is devoted to present a logical representation language for the formalization of expectation-based emotions. I will explain the syntax and the semantics of this logic which allows to represent different kinds of an agent's mental states such as knowledge, graded belief and graded goal. In Section 1.3, the logical framework of Section 1.2 will be applied to the formalization of the cognitive structures of hope and fear and of their intensities. I will consider the concepts of hope and fear. In Section 1.4 I will discuss some related works in the area of logical modeling of emotions. Then, I will conclude.

1.2 A simple logic of graded beliefs and graded goals

The logic LGA (*Logic of Graded mental Attitudes*) presented in this section is a BDI-like logic in the sense of [10, 32] which allows to represent formally different kinds of mental attitudes of an agent including knowledge, graded beliefs (*i.e.*, believing with a certain strength that a given proposition is true) and graded goals (*i.e.*, wanting with a given force or strength a given proposition to be true). I will present the syntax and the semantics of the logic LGA. I will only consider the single-agent case, postponing to future work an extension of the logic to the multi-agent case.

1.2.1 Syntax

Assume a finite set of propositional variables $Prop = \{p, q, \ldots\}$ and a finite set of positive integers $Num = \{0, \ldots, \mathsf{max}\}$ with $\mathsf{max} \geq 1$. The language $\mathcal{L}$ of the logic LGA is the set of formulae defined by the following grammar in Backus-Naur Form (BNF):

$$Atm : \chi ::= p \mid \mathsf{exc}_{\mathsf{h}} \mid \mathsf{des}_{\mathsf{h}}$$
$$Fml : \varphi ::= \chi \mid \neg\varphi \mid \varphi \wedge \varphi \mid \mathsf{K}\varphi$$

where p ranges over $Prop$ and h ranges over Num. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from p, $\neg$ and $\wedge$ in the standard way.

Atm is the set of atomic formulae. The latter includes propositional variables and special constructions which are used to represent the agent's mental state (the agent's beliefs and goals).

The special atoms $\mathsf{exc}_{\mathsf{h}}$ are used to identify the degree of *plausibility* of a given world according to the agent. Indeed, possible worlds are ordered according to their plausibility degree for the agent in such a way that the agent is capable of assessing whether a given world is more plausible than another world. Starting from [22], ranking among possible worlds have been extensively used in belief revision theory in order to define a selection mechanism (*i.e.*, a revision function) which can tell how to decide rationally which sentences to give up and which to keep when revising a knowledge base. I here use the notion of plausibility first introduced by Spohn [38]. Following Spohn's theory, the worlds that are assigned the smallest numbers are the most plausible, according to the beliefs of the individual. That is, the ordinal h assigned to a given world rather captures the degree of *exceptionality* of this world, where the exceptionality degree of a world is nothing but the opposite of its plausibility degree (*i.e.*, the exceptionality degree of a world decreases when its plausibility degree increases). Therefore, formula $\mathsf{exc}_{\mathsf{h}}$ can be read alternatively as "according to the agent, the current world has a degree of exceptionality h" or "the current world has a degree of plausibility $\mathsf{max} - \mathsf{h}$".

The special atoms $\mathsf{des}_{\mathsf{h}}$ are used to identify the degree of *desirability* (or the degree of *goodness*) of a given world for the agent.[2]

Contrary to plausibility, the worlds that are assigned the biggest numbers are the most desirable for the agent. The degree of undesirability (or degree of a badness) of a given world is the opposite of its desirability degree (or degree of goodness). Therefore, formula $\mathsf{des}_{\mathsf{h}}$ can be read alternatively as "the current world has a degree of desirability h" or "the current world has a degree of undesirability $\mathsf{max} - \mathsf{h}$".

The formula $\mathsf{K}\varphi$ has to be read "the agent knows that φ is true". This concept of knowledge is the standard S5-notion, partition-based and fully introspective, that is commonly used both in computer science [17] and economics [4]. If a proposition is part of the agent's knowledge then it means that the agent considers it a well-established truth [42]. The dual of the operator K is denoted by $\widehat{\mathsf{K}}$, that is, we define:

$$\widehat{\mathsf{K}}\varphi \stackrel{\mathsf{def}}{=} \neg\mathsf{K}\neg\varphi$$

$\mathsf{K}\varphi$ has to be read "the agent thinks that φ is possible" or "the agent envisages a situation in which φ is true".

1.2.2 Semantics

The model-theoretic semantics of the logic LGA is a possible world semantics.

[2] The term 'goodness' is perhaps more convenient because one may ascribe to the terms 'desire' and 'desirability' a hedonistic connotation that I do not want to use here.

Definition 1 (Model). LGA-models are tuples $M = \langle W, \sim, \kappa_{\mathsf{exc}}, \kappa_{\mathsf{des}}, \mathcal{V}\rangle$ where:

- W is a nonempty set of possible worlds or states;
- $\sim$ is an equivalence relation between worlds in W;
- $\kappa_{\mathsf{exc}} : W \longrightarrow Num$ and $\kappa_{\mathsf{des}} : W \longrightarrow Num$ are functions from the set of possible worlds into the set of integers Num;
- $\mathcal{V} : W \longrightarrow 2^{Prop}$ is a valuation function.

As usual, $p \in \mathcal{V}(w)$ means that proposition p is true at world w.

The accessibility relation $\sim$, which is used to interpret the epistemic operator K, can be viewed as a function from W to 2^W. Therefore, we can write $\sim(w) = \{v \in W : w \sim v\}$. The set $\sim(w)$ is the agent's *information state* at world w: the set of worlds that the agent considers possible at world w or, the set of worlds that the agent cannot distinguish from world w. As $\sim$ is an equivalence relation, if $w \sim v$ then the agent has the same information state at w and v (*i.e.*, the agent has the same knowledge at w and v).

The function κ_{exc} represents a plausibility grading of the possible worlds and is used to interpret the atomic formulae $\mathsf{exc}_{\mathsf{h}}$. $\kappa_{\mathsf{exc}}(w) = \mathsf{h}$ means that, according to the agent the world w has a degree of exceptionality h or, alternatively, according to the agent the world w has a degree of plausibility $\mathsf{max} - \mathsf{h}$. (Remember that the degree of exceptionality of a world is nothing but the opposite of its plausibility degree.) The function κ_{exc} allows to model the notion of belief: among the worlds the agent cannot distinguish from a given world w (*i.e.*, the agent's information state at w), there are worlds that the agent considers more plausible than others. For example, suppose that $\sim(w) = \{w, v, u\}$, $\kappa_{\mathsf{exc}}(w) = 2$, $\kappa_{\mathsf{exc}}(u) = 1$ and $\kappa_{\mathsf{exc}}(v) = 0$. This means that at world w the agent cannot distinguish the three worlds w, v and u (*i.e.*, $\{w, v, u\}$ is the set of worlds that the agent considers possible at world w). Moreover, according to the agent, the world v is strictly more plausible than the world u and the world u is strictly more plausible than the world w (as $\mathsf{max}-0 > \mathsf{max}-1 > \mathsf{max}-2$).

The function κ_{des} is used to interpret the atomic formulae $\mathsf{des}_{\mathsf{h}}$. $\kappa_{\mathsf{des}}(w) = \mathsf{h}$ means that, according to the agent the world w has a degree of desirability (or goodness) h or, alternatively, according to the agent the world w has a degree of undesirability (or badness) $\mathsf{max} - \mathsf{h}$. (Remember that the degree of undesirability of a world is the opposite of its desirability degree.)

LGA-models are supposed to satisfy the following additional *normality* constraints for the plausibility grading and for the desirability grading:

($NORM_{\kappa_{\mathsf{exc}}}$) according to the agent, there exists a world with maximal degree of plausibility (or with minimal degree of exceptionality): for every $w \in W$, there is v such that $w \sim v$ and $\kappa_{\mathsf{exc}}(v) = 0$;

($NORM_{\kappa_{\mathsf{des}}}$) according to the agent, there exists a world with minimal degree of desirability 0 (or with maximal degree of undesirability): for every $w \in W$, there is v such that $w \sim v$ and $\kappa_{\mathsf{des}}(v) = 0$.

As I will show in Section 1.2.4, these should be interpreted as a sort of *rationality* requirements that ensure that an agent cannot have inconsistent beliefs or conflicting goals.

Definition 2 (Truth conditions). Given a LGA-model M, a world w and a formula φ, $M, w \models \varphi$ means that φ is true at world w in M. The rules defining the truth conditions of atomic formulae, negation, conjunction and the epistemic operators are:

- $M, w \models p$ iff $p \in \mathcal{V}(w)$
- $M, w \models \mathsf{exc}_{\mathsf{h}}$ iff $\kappa_{\mathsf{exc}}(w) = \mathsf{h}$
- $M, w \models \mathsf{des}_{\mathsf{h}}$ iff $\kappa_{\mathsf{des}}(w) = \mathsf{h}$
- $M, w \models \neg\varphi$ iff not $M, w \models \varphi$
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$
- $M, w \models \mathsf{K}\varphi$ iff $M, v \models \varphi$ for all v with $w \sim v$

In the sequel I write $\models_{\mathsf{LGA}} \varphi$ to mean that φ is *valid* in LGA (φ is true in all LGA-models).

1.2.3 Definitions of graded belief, certain belief, goal and graded goal

Following [38], I extend the plausibility degree of a possible world to a plausibility degree of a formula viewed as a set of worlds (the worlds where the formula is satisfiable).

Definition 3 (Exceptionality degree of a formula). Let $||\varphi||_w = \{v \in W : M, v \models \varphi \text{ and } w \sim v\}$ be the set of worlds envisaged by the agent at world w in which φ is true. The exceptionality degree of a formula φ at world w, denoted by $\kappa^w_{\mathsf{exc}}(\varphi)$, is defined as follows:

$$\kappa^w_{\mathsf{exc}}(\varphi) = \begin{cases} \min\limits_{v \in ||\varphi||_w} \kappa_{\mathsf{exc}}(v) & \text{if } ||\varphi||_w \neq \emptyset \\ \mathsf{max} & \text{if } ||\varphi||_w = \emptyset \end{cases}$$

As expected, the *plausibility* degree of a formula φ is defined as $\mathsf{max} - \kappa^w_{\mathsf{exc}}(\varphi)$.

I do a similar manipulation for the desirability degree.

Definition 4 (Desirability degree of a formula). The desirability degree of a formula φ at world w, denoted by $\kappa^w_{\mathsf{des}}(\varphi)$, is defined as follows:

$$\kappa^w_{\mathsf{des}}(\varphi) = \begin{cases} \min\limits_{v \in ||\varphi||_w} \kappa_{\mathsf{des}}(v) & \text{if } ||\varphi||_w \neq \emptyset \\ 0 & \text{if } ||\varphi||_w = \emptyset \end{cases}$$

The *undesirability* degree of a formula φ is defined as $\mathsf{max} - \kappa^w_{\mathsf{des}}(\varphi)$.

Again following [38], I define semantically the concept of belief as a formula which is true in all worlds that are maximally plausible (or minimally exceptional) according to the agent.

Definition 5 (Belief, Bel). At world w the agent believes that φ is true, *i.e.*, $M, w \models \mathsf{Bel}\varphi$, if and only if, for every v such that $w \sim v$, if $\kappa_{\mathsf{exc}}(v) = 0$ then $M, v \models \varphi$.

The following concept of graded belief is taken from [24, 23]: the strength of the belief that φ is equal to the exceptionality degree of $\neg\varphi$.[3]

Definition 6 (Graded belief, $\mathsf{Bel}^{\mathsf{h}}$). At world w the agent believes that φ with strength equal to h, *i.e.*, $M, w \models \mathsf{Bel}^{\mathsf{h}}\varphi$, if and only if, $\kappa^{w}_{\mathsf{exc}}(\neg\varphi) = \mathsf{h}$.

I moreover define the following concept of certain belief in the sense of believing that φ is true with maximal strength max.

Definition 7 (Certain belief, $\mathsf{Certain}$). At world w the agent is certain that φ is true, *i.e.*, $M, w \models \mathsf{Certain}\varphi$, if and only if $\kappa^{w}_{\mathsf{exc}}(\neg\varphi) = \mathsf{max}$.

The following concept of graded goal is the motivational counterpart of the notion of graded belief. I say that at world w the agent wants (or wishes) φ to be true with strength equal to h if and only if, the desirability of φ is equal to h.

Definition 8 (Graded goal, $\mathsf{Goal}^{\mathsf{h}}$). At world w the agent wants/wishes φ to be true with strength equal to h (or the agent has the goal that φ with strength equal to h), *i.e.*, $\mathsf{Goal}^{\mathsf{h}}\varphi$, if and only if, $\kappa^{w}_{\mathsf{des}}(\varphi) = \mathsf{h}$.

Note that previous definition captures a 'pessimistic' notion of graded goal: when assessing how much φ is desirable, the agent focuses on his epistemic alternatives with a minimal degree of desirability in which φ is true. In this sense, the preceding notion of graded goal that φ consists in a kind of worst case analysis of the epistemic alternatives in which φ is true. It is also worth noting that the previous operator of graded goal is an operator of strong possibility (or actual possibility) in the sense of possibility theory [14]. In the context of possibility theory it is also called operator Δ.[4]

The reason why the definition of graded goal is not symmetric to the definition of graded belief is that these two concepts satisfy different logical properties. As I will show below in Section 1.2.4, Definition 6 and Definition 8 allow to capture interesting differences between graded belief and graded goal, in particular on the way they distribute over conjunction and over disjunction.

The notion of goal is a just a special case of the notion of graded goal. I say that the agent wants (or wishes) φ to be true with strength equal to h if and only if, the agent wants (or wishes) φ to be true with strength higher than 0.

Definition 9 (Goal, Goal). At world w the agent wants/wishes φ to be true (or the agent has the goal that φ), *i.e.*, $\mathsf{Goal}\varphi$, if and only if $\kappa^{w}_{\mathsf{des}}(\varphi) > 0$.

As the following proposition highlights the concepts of belief, graded belief, certain belief, goal and graded goal semantically defined in Definitions 5-9 are all syntactically expressible in the logic LGA.

Proposition 1. *For every model M, world w and $\mathsf{h} \in Num$:*

- $M, w \models \mathsf{Bel}\varphi$ *iff* $M, w \models \mathsf{K}(\mathsf{exc}_0 \rightarrow \varphi)$

[3] Similar operators for graded belief are studied in [3, 43].

[4] See also [1] for a recent application of the operator Δ to modeling desires.

- $M, w \models \mathsf{Bel}^{\mathsf{h}}\varphi$ *iff* $\begin{cases} M, w \models \widehat{\mathsf{K}}(\mathsf{exc}_{\mathsf{h}} \wedge \neg\varphi) \wedge \mathsf{K}(\mathsf{exc}_{<\mathsf{h}} \rightarrow \varphi) & \text{if } h < \mathsf{max} \\ M, w \models \mathsf{K}(\mathsf{exc}_{<\mathsf{h}} \rightarrow \varphi) & \text{if } h = \mathsf{max} \end{cases}$
- $M, w \models \mathsf{Certain}\varphi$ *iff* $M, w \models \mathsf{K}(\mathsf{exc}_{<\mathsf{max}-1} \rightarrow \varphi)$
- $M, w \models \mathsf{Goal}^{\mathsf{h}}\varphi$ *iff* $\begin{cases} M, w \models \widehat{\mathsf{K}}(\mathsf{des}_{\mathsf{h}} \wedge \varphi) \wedge \mathsf{K}(\mathsf{des}_{<\mathsf{h}} \rightarrow \neg\varphi) & \text{if } h < \mathsf{max} \\ M, w \models \mathsf{K}(\mathsf{des}_{<\mathsf{h}} \rightarrow \neg\varphi) & \text{if } h = \mathsf{max} \end{cases}$
- $M, w \models \mathsf{Goal}\varphi$ *iff* $M, w \models \mathsf{K}(\mathsf{des}_0 \rightarrow \neg\varphi)$

where $\mathsf{exc}_{<\mathsf{k}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{y} \in Num: 0 \leq \mathsf{y} < \mathsf{k}} \mathsf{exc}_{\mathsf{y}}$ *and* $\mathsf{des}_{<\mathsf{k}} \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{y} \in Num: 0 \leq \mathsf{y} < \mathsf{k}} \mathsf{des}_{\mathsf{y}}$ *for all* $\mathsf{k} \in Num$ *such that* $\mathsf{k} \geq 1$, $\mathsf{exc}_{<0} \stackrel{\mathtt{def}}{=} \bot$ *and* $\mathsf{des}_{<0} \stackrel{\mathtt{def}}{=} \bot$.

1.2.4 Some properties of mental attitudes

The following are some examples of LGA-validity which capture the basic relationships between knowledge, belief, graded belief, graded goal and certain belief. For every $\mathsf{h} \in Num$ we have:

$$\models_{\mathsf{LGA}} \mathsf{Bel}^{\mathsf{h}}\varphi \rightarrow \mathsf{Bel}\varphi \text{ if } \mathsf{h} > 0 \tag{1.1}$$

$$\models_{\mathsf{LGA}} \mathsf{Certain}\varphi \leftrightarrow \mathsf{Bel}^{\mathsf{max}}\varphi \tag{1.2}$$

$$\models_{\mathsf{LGA}} \mathsf{Goal}\varphi \leftrightarrow \bigvee_{\mathsf{h} \in Num: \mathsf{h} > 0} \mathsf{Goal}^{\mathsf{h}}\varphi \tag{1.3}$$

$$\models_{\mathsf{LGA}} \neg(\mathsf{Bel}\varphi \wedge \mathsf{Bel}\neg\varphi) \tag{1.4}$$

$$\models_{\mathsf{LGA}} \neg(\mathsf{Goal}\varphi \wedge \mathsf{Goal}\neg\varphi) \tag{1.5}$$

According to the validity 1.1, believing φ is the same as believing φ with degree 1. According to the validity 1.2, being certain that φ is true is the same as believing φ with maximal degree max. Finally, according to the validity 1.3, having the goal that φ is true is the same as having a goal that φ is true with some strength higher than 0. Validities 1.4 and 1.5 are consequences of the constraints $(NORM_{\kappa_{\mathsf{exc}}})$ and $(NORM_{\kappa_{\mathsf{des}}})$ given in Section 1.2.2. Their meaning is that agent are assumed to be rational, in the sense that they cannot have at the same time logically inconsistent beliefs or logically inconsistent goals.

The following four valid formulae capture the basic decomposability properties of the operators of graded belief and graded goal:

$$\models_{\mathsf{LGA}} (\mathsf{Bel}^{\mathsf{h}}\varphi \wedge \mathsf{Bel}^{\mathsf{k}}\psi) \rightarrow \mathsf{Bel}^{\geq \max\{\mathsf{h},\mathsf{k}\}}(\varphi \vee \psi) \tag{1.6}$$

$$\models_{\mathsf{LGA}} (\mathsf{Goal}^{\mathsf{h}}\varphi \wedge \mathsf{Goal}^{\mathsf{k}}\psi) \rightarrow \mathsf{Goal}^{\min\{\mathsf{h},\mathsf{k}\}}(\varphi \vee \psi) \tag{1.7}$$

$$\models_{\mathsf{LGA}} (\mathsf{Bel}^{\mathsf{h}}\varphi \wedge \mathsf{Bel}^{\mathsf{k}}\psi) \rightarrow \mathsf{Bel}^{\min\{\mathsf{h},\mathsf{k}\}}(\varphi \wedge \psi) \tag{1.8}$$

$$\models_{\mathsf{LGA}} (\mathsf{Goal}^{\mathsf{h}}\varphi \wedge \mathsf{Goal}^{\mathsf{k}}\psi) \rightarrow \mathsf{Goal}^{\geq \max\{\mathsf{h},\mathsf{k}\}}(\varphi \wedge \psi) \tag{1.9}$$

where:

$$X^{\geq \mathsf{h}}\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{k} \in Num: \mathsf{k} \geq \mathsf{h}} X^{\mathsf{k}}\varphi \text{ with } X \in \{\mathsf{Bel}, \mathsf{Goal}\}$$

According to the validity 1.6, the degree of belief of $\varphi \vee \psi$ is at least equal to the maximum of the degree of belief of φ and ψ. According to the validity 1.7, the strength of the goal that $\varphi \vee \psi$ is equal to the minimum of the strength of the goal of φ and ψ. According to the validity 1.8, the degree of belief of $\varphi \wedge \psi$ is equal to the minimum of the degree of belief of φ and ψ. According to the validity 1.9, the strength of the goal that $\varphi \wedge \psi$ is at least equal to the maximum of the strength of the goal of φ and ψ.

The interesting aspect of the preceding valid formulae is that graded goals distribute over conjunction and over disjunction in the opposite way as graded beliefs. Consider for instance the validity (1.8) and compare it to the validity (1.9). The joint occurrence of two events φ and ψ cannot be more plausible than the occurrence of a single event. This is the reason why in the right side of the validity (1.8) we have the min. On the contrary, the joint occurrence of two desirable events φ and ψ is more desirable than the occurrence of a single event. This is the reason why in the right side of the validity (1.9) we have the $\geq$ max. For example, suppose Peter wishes to go to the cinema in the evening with strength h (*i.e.*, $\mathsf{Goal}^{\mathsf{h}} goToCinema$) and, at the same time, he wishes to spend the evening with his girlfriend with strength k (*i.e.*, $\mathsf{Goal}^{\mathsf{k}} stayWithGirlfriend$). Then, according to the validity (1.9), Peter wishes to to go the cinema with his girlfriend with strength at least $\max\{\mathsf{h},\mathsf{k}\}$ (*i.e.*, $\mathsf{Goal}^{\geq \max\{\mathsf{h},\mathsf{k}\}}(goToCinema \wedge stayWithGirlfriend)$).

Note that the following formula is valid in the logic LGA:

$$\models_{\mathsf{LGA}} \mathsf{K}\neg\varphi \rightarrow \mathsf{Goal}^{\mathsf{max}}\varphi \tag{1.10}$$

This means that if in all situations that an agent envisages φ is false, then the agent wants φ to be true with maximal strength max. This property follows from the fact that the graded goal operator $\mathsf{Goal}^{\mathsf{h}}$ represents a 'pessimistic' notion of goal, that is to say, it is assumed that when assessing how much φ is desirable, the agent focuses on his epistemic alternatives with a minimal degree of desirability in which φ is true. Therefore, if the agent does not envisage a situation in which φ is true, he will consider φ maximally desirable. Nonetheless one might find the preceding property counterintuitive. For this reason, I define the following concept of *realistic goal*, which does not satisfy the same property:

$$\mathsf{RGoal}^{\mathsf{h}}\varphi \stackrel{\mathsf{def}}{=} \mathsf{Goal}^{\mathsf{h}}\varphi \wedge \widehat{\mathsf{K}}\varphi$$

This means an agent has a realistic goal that φ of strength h, denoted by $\mathsf{RGoal}^{\mathsf{h}}\varphi$, if and only if the agent wants φ to be true with strength h and envisages at least one situation in which φ is true. Obviously if the agent does not envisage a situation in which φ is true then he does not have φ as a realistic goal. That is, for any $\mathsf{h} \in Num$:

$$\models_{\mathsf{LGA}} \mathsf{K}\neg\varphi \rightarrow \neg\mathsf{RGoal}^{\mathsf{h}}\varphi \tag{1.11}$$

1.2.5 The concept of goal: some remarks

In this work I assume that an agent's goal has two dimensions: (1) its content, and (2) its strength or value. Following Castelfranchi's theory of goals [6, 11], with the term 'value of a goal' I mean the subjective importance of the goal for the agent strictly dependent on context conditions and mental attitudes. In the graded goal formula $\mathsf{Goal}^{\mathsf{h}}\varphi$, φ is the goal content while h is the goal strength (or goal value).

Most appraisal models of emotions (see, *e.g.*, [25, 36]) assume that explicit evaluations based on evaluative beliefs (*i.e.*, the belief that a certain event is desirable or undesirable, good or bad, pleasant or unpleasant) are a necessary constituent of emotional experience. Other appraisal models (see, *e.g.*, [37, 19, 21, 9]) assume that emotions are triggering by specific combinations of beliefs and goals (or desires), and that the link between cognition and emotion is not necessarily mediated by evaluative beliefs. Reisenzein [34] calls *cognitive-evaluative* the former and *cognitive-motivational* the latter kind of models. For example, according to cognitive-motivational models of emotions, a person's happiness about a certain fact φ can be reduced to the person's belief that φ obtains and the person's desire (or goal) that φ obtains. On the contrary, according to cognitive-evaluative models, a person feels happy about a certain fact φ if she believes that φ obtains and she evaluates φ to be desirable (or good) for her. A similar distinction has been discussed in philosophy on whether motivational mental states such as goals and desires are derived from and reduced to evaluative beliefs or, viceversa, whether evaluative beliefs are derived from goals or desires (*i.e.*, are desires and goals more primitive than evaluative beliefs, or viceversa?) (see, *e.g.*, [26, 27, 5]).

In the present work, I stay closer to cognitive-evaluative models. In fact, I reduce the notion of goal - involved in the appraisal configuration of a given emotion - to the agent's evaluation of the desirability (or goodness) of a given state of affairs.[5] That is, I assume that an agent has φ as a goal *if and only if* the situations envisaged by the agent in which φ is true are desirable for him. Consequently, I assume that an agent's positive/negative emotion requires the agent's evaluation of the desirability of a certain event, situation or object. One might argue that evaluative beliefs are not primitive mental states, but they are just derived from goals. That is, an agent evaluates a given situation or event to be desirable (or good) for him, *because* he believes that in this situation he will achieve his goals or *because* he believes that the event has positive implications on his goals. I here take a different point of view by assuming that evaluations and goals somehow coincide.

[5] Note that, in most cases, the assessment of the desirability of φ has a somatic component, that is, the agent considers φ desirable because while thinking and imagining a situation in which φ holds, the agent has a positive *feeling* or *sensation*.

1.3 Formalization of expectation-based emotions and their intensity

The modal operators of graded belief and graded goal defined in Section 1.2.2 are used here to provide a logical analysis of expectation-based emotions such as hope and fear and of their intensities. An expectation-based emotion is an emotion that an agent experiences when having either a positive or a negative expectation about a certain fact φ, that is, when believing that φ is true with a certain strength but envisaging the possibility that φ could be false and either (1) having the goal that φ is true (positive expectation) or (2) having the goal that φ is false (negative expectation).

According to some psychological models [34, 35, 25, 31] and computational models [20, 15] of emotions, the intensity of hope with respect to a given event is a monotonically increasing function of: the degree to which the event is desirable and the likelihood of the event. That is, the higher is the desirability of the event φ, and the higher is the intensity of the agent's hope that this event will occur; the higher is the likelihood of the event φ, and the higher is the intensity of the agent's hope that this event will occur.[6] Analogously, the intensity of fear with respect to a given event is a monotonically increasing function of: the degree to which the event is undesirable and the likelihood of the event. That is, the higher is the undesirability of the event φ, and the higher is the intensity of the agent's fear that this event will occur; the higher is the likelihood of the event φ, and the higher is the intensity of the agent's fear that this event will occur. There are several possible merging functions which satisfy these properties. For example, I could define the merging function *merge* for calculating emotion intensity as an average function, according to which the intensity of hope about a certain event φ is the average of the strength of the belief that φ will occur and the strength of the goal that φ will occur. That is, for every $\mathsf{h}, \mathsf{k} \in \mathit{Num}$ representing respectively the strength of the belief and the strength of the goal, I could define $\mathit{merge}(\mathsf{h},\mathsf{k})$ as[7]

$$\begin{cases} \frac{\mathsf{h}+\mathsf{k}}{2} & \text{if } \mathsf{h} > 0 \text{ and } \mathsf{h} > 0 \\ 0 & \text{if } \mathsf{h} = 0 \text{ or } \mathsf{h} = 0 \end{cases} \tag{1.12}$$

Another possibility is to define *merge* as a product function (also used in [20]), according to which the intensity of hope about a certain event φ is the product of the strength of the belief that φ will occur and the strength of the goal that φ will occur. That is, for every $\mathsf{h}, \mathsf{k} \in \mathit{Num}$ I could define $\mathit{merge}(\mathsf{h},\mathsf{k})$ as

$$\mathsf{h} \times \mathsf{k} \tag{1.13}$$

[6] According to Ortony et al. [31] the intensity of hope and fear is determined by a third parameter: the (temporal and spatial) *proximity* to the expected event (the higher is the proximity to the expected event, and the higher is the intensity of hope/fear). This third dimension is not considered in the present analysis.

[7] The second condition is necessary to ensure that intensity of emotion is equal to zero when one of the two parameters (belief strength or goal strength) is set to zero.

Here I do not choose a specific merging function, as such choice would require an experimental validation and would much depend on the domain of application in which the formal model has to be used.

Let me now define the notion of hope and fear with their corresponding intensities. As pointed out in the introduction of the paper, I only characterize the emotion's triggering conditions, that is, the agent's mental states (beliefs and goals) that are responsible for triggering the agent's emotional reaction and that 'cause' the agent to feel the emotion. I define

$$ISCALE = \{\mathsf{x}: \text{ there are } \mathsf{h}, \mathsf{k} \in Num \text{ such that } merge(\mathsf{h},\mathsf{k}) = \mathsf{x}\}$$

to be the emotion intensity scale (*i.e.*, the set of values over which the intensity of hope or fear can range).

An agent is experiencing a hope with with intensity x about φ if and only if there are $\mathsf{h}, \mathsf{k} \in Num$ such that $\mathsf{h} < \mathsf{max}$, h is the strength to which the agent believes that φ is true, k is the strength to which the agent realistically wants φ to be true and $\mathsf{x} = merge(\mathsf{h},\mathsf{k})$. That is:

$$\mathsf{Hope}^{\mathsf{x}}\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{h},\mathsf{k} \in Num: \mathsf{h}<\mathsf{max} \text{ and } merge(\mathsf{h},\mathsf{k})=\mathsf{x}} (\mathsf{Bel}^{\mathsf{h}}\varphi \wedge \mathsf{RGoal}^{\mathsf{k}}\varphi)$$

The notion of fear can be defined in a similar way, after assuming that an event φ is undesirable for the agent if and only if the agent wants φ to be false.[8] An agent is experiencing a fear with with intensity x about φ if and only if there are $\mathsf{h}, \mathsf{k} \in Num$ such that $\mathsf{h} < \mathsf{max}$, h is the strength to which the agent believes that φ is true, k is the strength to which the agent realistically wants φ to be false and $\mathsf{x} = merge(\mathsf{h},\mathsf{k})$. That is:

$$\mathsf{Fear}^{\mathsf{x}}\varphi \stackrel{\mathtt{def}}{=} \bigvee_{\mathsf{h},\mathsf{k} \in Num: \mathsf{h}<\mathsf{max} \text{ and } merge(\mathsf{h},\mathsf{k})=\mathsf{x}} (\mathsf{Bel}^{\mathsf{h}}\varphi \wedge \mathsf{RGoal}^{\mathsf{k}}\neg\varphi)$$

The reason why in the definitions of hope and fear I use the notion of *realistic goal* $\mathsf{RGoal}^{\mathsf{k}}$ instead of a *simple goal* $\mathsf{Goal}^{\mathsf{k}}$ is that I want to avoid that if an agent does not envisage a situation in which φ is true (*i.e.*, $\mathsf{K}\neg\varphi$) then he necessarily feels fearful about $\neg\varphi$.[9]

Moreover, in the preceding definitions, the strength of the belief is supposed to be less than max in order distinguish emotions such as hope and fear implying some form of uncertainty from emotions such as happiness and sadness (or unhappiness) which are based on certainty. In order to experience hope (or fear) about φ, the agent should have a minimal degree of uncertainty that φ might be false (*i.e.*, the agent should not know that φ is true). Indeed, from

[8] I am aware that this is a simplifying assumption, as the undesirability of an event φ does not always coincide with the desirability of its negation. For example, an agent might desire 'to gain 100 €', even though 'not gaining 100 €' is not undesirable for him (the agent is simply indifferent about this result).

[9] In fact, $\mathsf{K}\neg\varphi$ implies both $\mathsf{Goal}^{\mathsf{max}}\varphi$ (see the equation 1.10 in Section 1.2.4.) and $\mathsf{Bel}\neg\varphi$. Therefore, if I had used $\mathsf{Goal}^{\mathsf{k}}$ in the preceding definition of fear, $\mathsf{K}\neg\varphi$ would have implied $\mathsf{Fear}^{\mathsf{x}}\neg\varphi$ for some number $\mathsf{x} \in ISCALE$ which is counterintuitive.

the previous definitions, it follows that:

$$\models_{\mathsf{LGA}} \mathsf{Hope}^{\mathsf{x}}\varphi \rightarrow \neg\mathsf{Certain}\varphi \tag{1.14}$$

$$\models_{\mathsf{LGA}} \mathsf{Fear}^{\mathsf{x}}\varphi \rightarrow \neg\mathsf{Certain}\varphi \tag{1.15}$$

This means that if an agent hopes (resp. fears) φ to be true, then he is not certain that φ (*i.e.*, he does not have the strong belief that φ). For example, if I hope that my paper will be accepted for publication in a prestigious journal, then it means that I am not certain that my paper will be accepted. The preceding two validities are consistent with Spinoza's quote "Fear cannot be without hope nor hope without fear". Indeed, if an agent hopes that φ will be true then, according to the validity 1.14, he envisages the possibility that φ will be false. Therefore, he experiences some fear that φ will be false. Conversely, if an agent fears that φ will be true then, according to the validity 1.15, he envisages the possibility that φ will be false. Therefore, he experiences some hope that φ will be false.

Remark 1. It has to be noted that hope and fear, and more generally expectations, are not necessarily about a *future* state of affairs, but they can also be about a *present* state of affairs or a *past* state of affairs. For example, I might say 'I hope that you feel better now!' or 'I fear that you did not enjoy the party yesterday night!'.

On the contrary, to feel happy (resp. sad) about φ, the agent should be *certain* that φ is true. For example, if I am happy that my paper has been accepted for publication in a prestigious journal, then it means that I am certain that my paper has been accepted. More precisely, an agent is experiencing happiness with intensity h about φ if and only if, the agent strong believes (is certain) that φ is true and h is the strength to which the agent wants φ to be true. That is:

$$\mathsf{Happiness}^{\mathsf{h}}\varphi \stackrel{\mathtt{def}}{=} \mathsf{Certain}\varphi \wedge \mathsf{RGoal}^{\mathsf{h}}\varphi$$

Moreover, an agent is experiencing sadness with intensity h about φ if and only if, the agent strongly believes (is certain) that φ is true and h is the strength to which the agent wants φ to be false. That is:

$$\mathsf{Sadness}^{\mathsf{h}}\varphi \stackrel{\mathtt{def}}{=} \mathsf{Certain}\varphi \wedge \mathsf{RGoal}^{\mathsf{h}}\neg\varphi$$

1.4 Related works

Emotion is a very active field not only in psychology but also in AI. Several computational architectures of affective agents have been proposed in the last few years (see, *e.g.*, [33, 16, 13, 15]). The cognitive architecture EMA (Emotion and Adaption) [20] is one of the best example of research in this area. EMA defines a domain independent taxonomy of appraisal variables stressing the many different relations between emotions and cognition, by enabling a wide

range of internal appraisal and coping processes used for reinterpretation, shift of motivations, goal reconsideration *etc.*

There are also several researchers who have developed formal languages for reasoning about emotions and for modelling affective agents. I discuss here some of the most important formal approaches to emotions and compare them with the approach presented in this paper.

One of the most prominent logical analysis of emotions is the one proposed by Meyer and coll. [29, 39, 41]. In order to formalize emotions, they exploit the logical framework KARO [30]: a framework based on a blend of dynamic logic with epistemic logic, enriched with modal operators for motivational attitudes such as desires and goals. In Meyer et al.'s approach each instance of emotion is represented with a special predicate, or fluent, in the jargon of reasoning about action and change, to indicate that these predicates change over time. For every fluent a set of effects of the corresponding emotions on the agent's planning strategies are specified, as well as the preconditions for triggering the emotion. The latter correspond to generation rules for emotions. For instance, in [29] generation rules for four basic emotions are given: joy, sadness, anger and fear, depending on the agent's plans. More recently [41], generation rules for social emotions such as guilt and shame have been proposed.

Contrarily to Meyer et al.'s approach, in the logic LGA there are no specific formal constructs, like special predicates or fluents, which are used to denote that a certain emotion arises at a certain time. I just *define* the appraisal pattern of a given emotion in terms of some cognitive constituents such as goal and knowledge. For instance, according to the definition of hope proposed in Section 1.3, an agent experiences hope about φ if and only if, he believes φ and he wants φ to be true with a certain strength. In other words, following the so-called appraisal theories in psychology, in this work I only consider the appraisal variables of emotion which can be defined through the basic concepts of a BDI logic (*e.g.*, knowledge, belief, desire, goal).

In a more recent work [40] Meyer et al. have integrated quantitative aspects in their logical model of emotions. However, differently from the present approach, they do not study the variables determining intensities but instead focus on the integration of intensities into a qualitative model of emotions. For example, they propose a function describing how the intensity of an emotion decreases over time.

Lorini & Schwarzentruber [28] have recently proposed a logical model of counterfactual emotions such regret and guilt, *i.e.*, those emotions based on counterfactual thinking about agents' choices and actions. Adam et al. [2] have exploited a BDI logic in order to provide a logical formalization of the emotion types defined in Ortony, Clore and Collins's model (OCC model) [31]. Similar to the approach presented in this paper, in Lorini & Schwarzentruber's approach and in Adam et al.'s approach emotion types are defined in terms of some primitive concepts (and corresponding modal operators) such as the concepts of belief, desire, action and responsibility which allow to capture the different appraisal variables of emotions. However, Lorini & Schwarzentruber's approach

and Adam et al.'s approach are purely qualitative and do not consider emotion intensity.

1.5 Conclusion

The logical analysis of expectation-based emotions presented in this paper is obviously very simplistic. It misses a lot of important psychological aspects. For example, the fact that mental states on which emotions such as hope and fear are based are usually joined with bodily activation and components, and these components shape the whole subjective state of the agent and determine his action tendencies [18].[10] I have only focused on the cognitive structure of emotions, without considering the *felt* aspect of emotions. This is of course a limitation of the model presented in this paper, as the intensity of emotion also depends on the presence of these somatic components (*e.g.*, the intensity of fear is amplified by the fact that, when experiencing this emotion I feel my stomach contracted, my throat dry, *etc.*)

However, an analysis of the cognitive structure of emotions (*i.e.*, identifying the mental states which determine a given type of emotion), as the one presented in this paper, is a necessary step for having an adequate understanding of affective phenomena. I postpone to future work a logical analysis of the basic relationships between emotion and action (*i.e.*, how an emotion with a given intensity determines the agent's future reactions),[11] and of the relationships between cognitive structure and somatic aspects of emotions (*i.e.*, how somatic components affect emotion intensity).

References

1. L. Godo A. Casali and C. Sierra. A graded BDI agent model to represent and reason about preferences. *Artificial Intelligence*, 175:1468–1478, 2012.
2. C. Adam, A. Herzig, and D. Longin. A logical formalization of the OCC theory of emotions. *Synthese*, 168(2):201–248, 2009.
3. G. Aucher. A combined system for update logic and belief revision. In *Proceedings of PRIMA 2004*, volume 3371 of *LNAI*, pages 1–18. Springer-Verlag, 2005.
4. R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.
5. R. Bradley and C. List. Desire as belief revisited. *Analysis*, 69(1):31–37, 2009.
6. C. Castelfranchi. Mind as an anticipatory device: For a theory of expectations. In M. De Gregorio, V. Di Maio, M. Frucci, and C. Musio, editors, *Proceedings of the First*

[10] According to Frijda, actions tendencies are [18, pp. 75] "...states of readiness to achieve or maintain a given kind of relationship with the environment. They can be conceived of as plans or programs to achieve such ends, which are put in a state of readiness." For example, the action tendency associated to fear is escape. According to Lazarus [25], there is an important difference between action tendencies and coping strategies. While the former are innately programmed unconscious reflexes and routines, the latter are the product of a conscious deliberation process.

[11] A first step in this direction has been taken in [12].

International Symposium on Brain, Vision, and Artificial Intelligence (BVAI 2005), volume 3704 of *LNCS*. Springer, 2005.
7. C. Castelfranchi and E. Lorini. Cognitive anatomy and functions of expectations. In F. Schmalhofer, R. M. Young, and G. Katz, editors, *Proceedings of the First European Cognitive Science Conference (EuroCogSci 2003)*, pages 377–379, Mahwah, NJ, 2003. Lawrence Erlbaum Associates.
8. C. Castelfranchi and E. Lorini. Cognitive anatomy and functions of expectations. In R. Sun, editor, *Proceedings of IJCAI'03 Workshop on Cognitive modelling of Agents and Multi-Agent Interactions*, pages 29–36, 2003.
9. C. Castelfranchi and M. Miceli. The cognitive-motivational compound of emotional experience. *Emotion Review*, 1(3):223–231, 2009.
10. P. R. Cohen and H. J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.
11. R. Conte and C. Castelfranchi. *Cognitive and social action*. London Univ. College of London Press, London, 1995.
12. M. Dastani and E. Lorini. A logic of emotions: from appraisal to coping. In *Proceedings of the Eleventh International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1133–1140 . ACM Press, 2012.
13. F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. D. Carolis. 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, 2003.
14. D. Dubois and H. Prade. Possibility theory: qualitative and quantitative aspects. In D. Gabbay and P. Smets, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, volume Quantified Representation of Uncertainty and Imprecision, volume 1, pages 169–226. Kluwer, 1998.
15. M. S. El-Nasr, J. Yen, and T. R. Ioerger. FLAME: Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multi-Agent Systems*, 3(3):219–257, 2000.
16. C. Elliot. *The Affective reasoner: A process model for emotions in a multi-agent system*. PhD thesis, Northwestern University, Institute for Learning Sciences, 1992.
17. R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.
18. N. Frijda. *The emotions*. Cambridge University Press, 1986.
19. R. M. Gordon. *The structure of emotions*. Cambridge University Press, Cambridge, 1987.
20. J. Gratch and S. Marsella. A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.
21. O. H. Green. *The emotions*. Cambridge University Press, Cambridge, 1992.
22. A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17:157–170, 1988.
23. N. Laverny and J. Lang. From knowledge-based programs to graded belief-based programs, part i: on-line reasoning. In *Proceedings of the Sixteenth European Conference on Artificial Intelligence (ECAI'04)*, pages 368–372. IOS Press, 2004.
24. N. Laverny and J. Lang. From knowledge-based programs to graded belief-based programs, part ii: off-line reasoning. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 497–502. Professional Book Center, 2005.
25. R. S. Lazarus. *Emotion and adaptation*. Oxford Univ. Press, New York, 1991.
26. D. Lewis. Desire as belief. *Mind*, 97:323–332, 1988.
27. D. Lewis. Desire as belief II. *Mind*, 105:303–313, 1996.
28. E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4):814–847, 2011.
29. J.-J. Ch. Meyer. Reasoning about emotional agents. *International Journal of Intelligent Systems*, 21(6):601–619, 2006.
30. J.-J. Ch. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40, 1999.
31. Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

32. A. S. Rao and M. Georgeff. Decision procedures for BDI logics. *Journal of Logic and Computation*, 8(3):293–344, 1998.
33. W. S. Reilly and J. Bates. Building emotional agents. Technical report, CMUCS -92-143, School of Computer science, Canergie Mellon University, 1992.
34. R. Reisenzein. Emotional experience in the computational belief-desire theory of emotion. *Emotion Review*, 1(3):214–222, 2009.
35. R. Reisenzein. Emotions as metarepresentational states of mind: naturalizing the belief-desire theory of emotion. *Cognitive Systems Research*, 10:6–20, 2009.
36. K. Scherer. Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, and T. Johnstone, editors, *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, Oxford, 2001.
37. J. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, 1983.
38. W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W. L. Harper and B. Skyrms, editors, *Causation in decision, belief change and statistics*, pages 105–134. Kluwer, 1998.
39. B. R. Steunebrink, M. Dastani, and J.-J. Ch. Meyer. A logic of emotions for intelligent agents. In *Proceedings of the 22th AAAI conference on Artificial Intelligence (AAAI'07)*, pages 142–147. AAAI Press, 2007.
40. B. R. Steunebrink, M. Dastani, and J.-J. Ch Meyer. A formal model of emotions: integrating qualitative and quantitative aspects. In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 256–260. IOS Press, 2008.
41. P. Turrini, J.-J. Ch. Meyer, and C. Castelfranchi. Coping with shame and sense of guilt: a dynamic logic account. *Journal of Autonomous Agents and Multi-Agent Systems*, 20(3):401–420, 2009.
42. J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17:129–155, 2007.
43. H. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.