# Affective Computing and Interaction:

## Psychological, Cognitive and Neuroscientific Perspectives

Didem Gökçay
*Middle East Technical University, Turkey*

Gülsen Yildirim
*Middle East Technical University, Turkey*

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Book Publications: Julia Mosemann
Acquisitions Editor: Lindsay Johnston
Development Editor: Dave DeRicco
Publishing Assistant: Milan Vracarich Jr.
Typesetter: Milan Vracarich Jr., Casey Conapitski
Production Editor: Jamie Snavely
Cover Design: Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com



Library of Congress Cataloging-in-Publication Data

Affective computing and interaction : psychological, cognitive, and neuroscientific perspectives / Didem Gokcay and Gulsen Yildirim, editors.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-61692-892-6 (hardcover) -- ISBN 978-1-61692-894-0 (ebook) 1. Human-computer interaction. 2. Human-machine systems. 3. Affect (Psychology)--Computer simulation. I. Gokcay, Didem, 1966- II. Yildirim, Gulsen, 1978-
QA76.9.H85A485 2011
004.01'9--dc22
2010041639

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

# Chapter 6
# For a 'Cognitive Anatomy' of Human Emotions and a Mind-Reading Based Affective Interaction

**Cristiano Castelfranchi**
*CNR, Italy*

## ABSTRACT

*Human emotions are based on typical configurations of beliefs, goals, expectations etc. In order to understand the complexity of affective processing in humans, reactions to stimuli, perception of our bodily reaction to events or just the feeling related to something should be considered but this is not adequate. Besides, our body does not respond just to external stimuli (events); it reacts to our interpretation of the stimulus, to the meaning of the event as well. In order to build affective architectures we also have to model the body, and its perception. In this chapter, with the help of these facts, the author will analyze the cognitive anatomies of simple anticipation-based emotions in addition to some complex social emotions.*

## INTRODUCTION

In this chapter, we present, in a rather synthetic way and without the possibility of extensively discussing the literature:

a. An explicit and analytical cognitive modeling of human emotions (cognitive 'anatomies' in terms of beliefs, goals, etc.);
b. The limits of this fundamental approach, and the need for its embodiment: modeling and integrating the bodily motions and signals, and what we feel;
c. Its application to computational models, artificial intelligence (AI), and human-computer interaction (HCI).

DOI: 10.4018/978-1-61692-892-6.ch006

The effects of complex emotions processed by humans go beyond reacting to stimuli, perceiving our bodily reaction to events, or feeling something. Especially complex human emotions are based on specific mental states; they are typical configurations of beliefs, goals, motives, expectations etc. In this chapter, we will analyze some typical mental configurations needed for (i) rather simple anticipation-based emotions ('hope', 'fear', 'disappointment', 'relief', 'joy') and (ii) complex social emotions like 'shame', 'envy', 'guilt', 'pity': their ingredients and their coherent structure. In particular, we will analyze shame and guilt in a very synthetic way.

We are in favor of a componential analysis of emotions (and in general, mental states and processes like 'expectation', 'need', 'trust', 'argument', etc.). This allows a systematic explicit model of the relationships within and among the substances to be modeled. However, one should also care about accounting for the unitary character of the mental-behavioral phenomena. On one side, being atomically decomposable, the complex mental states have their own emergent, specific, non-reducible properties and functions on the other side.

Our body does not respond to external stimuli or events based on pattern matching; it also reacts to our interpretation of the stimulus, to the meaning of the event; that is to a mental representation. In addition, the body reacts to merely endogenous representations, to mental events (like a counterfactual imagination). For example, it is always a thought that makes us blush. Of course for a complete real emotion, bodily activation and perception is necessary: at least in terms of the activation of the central memory trace of the bodily reaction (somatic marker), the evocation of some sensation. We feel our bodily response, but we ascribe it to that event or idea; this combination gives an emotional nature to both sides.

## EXPECTATIONS AND RELATED EMOTIONS

### Expectations vs. Predictions

'Expectations' are not just 'Predictions'; they are not fully synonyms. Therefore, we do not want to use expectations (like in the literature) just to mean predictions, that is, epistemic representations about the future. We consider, in particular, a 'forecast' as a mere belief about a future state of the world and we distinguish it from a simple hypothesis. The difference is in terms of degree of certainty: a hypothesis may involve the belief that future p is possible while a forecast has the belief that future p is probable. A forecast implies that the chance threshold has been exceeded.

Putting aside the degree of confidence (a general term for covering weak and strong predictions), for us expectations have a more restricted meaning (and this is why a computer can produce predictions or forecasts but do not have expectations). In 'expectations':

i. the prediction is relevant for the predictor; he is concerned, interested, and that is why
ii. he is expecting, that is the prediction is aimed at being verified; he is waiting in order to know whether the prediction is true or not.

Expectation is a suspended state after the formulation of a prediction[1]. If there is an expectation then there is a prediction, but not the other way around.

### Epistemic Goals and Activity

In the prediction-expectation chain, first of all, the agent X, has the Goal of knowing whether the predicted event or state really happens (epistemic goal). She is waiting for this; at least for curiosity. This concept of 'waiting for' and 'looking for' is necessarily related to the notion of expecting and expectation, but not to the notion of prediction.

During the expectation process, either X is actively monitoring what is happening and comparing the incoming information (for example perception) to the internal mental representation; or X is doing this cyclically and regularly; or X will compare what happens with her prediction (epistemic actions) in any case at the moment of the future event or state. Because in any case she has the Goal to know whether the world actually is as anticipated and if the prediction was correct. Therefore, in order to represent 'expecting' and the true 'expectation' schematically, we can write:

$$Expectation(x,p) \Rightarrow Bel(x,p^{t''})^{t'} \& Goal(x,p)$$

where $Bel(x,p^{t''})^{t'}$ is the belief of x at $t'$ that the predicted event p occurs at $t''$ (where $t'' > t'$) and $Goal(x,p)$ denotes the goal of x from $t'$ to $t'''$ ($t''' > t''$) for $Know(x,p^{t''}) \vee Know(x,\sim p^{t''})$.

## Content Goals

This Epistemic/Monitoring Goal is combined with Goal that p: the agent's need, desire, or 'intention that' the world should realize. The Goal that p is true (or the Goal that Not p). This is really why and in which sense X is concerned and not indifferent, and also why she is monitoring the world. She is an agent with interests, desires, needs, objectives, on the world, not just a predictor. This is also why computers, that already make predictions, do not have expectations.

Expectations can be classified according to the relation between the goals and predictions. When the agent has a goal opposite to her prediction, she has a negative expectation. On the other hand, when the agent has a goal equal to her prediction she has a positive expectation.

In sum, expectations are axiological anticipatory mental representations, endowed with Valence: they are positive or negative or ambivalent or neutral. But in any case they are evaluated against some concern, drive, motive, goal of the agent. In expectations, we have to distinguish two components:

- On one side, there is a mental anticipatory representation, the belief about a future state or event, the mental anticipation of the fact, what we might also call pre-vision (to foresee).
  The format of this belief or pre-vision can be either propositional or imagery (or mental model of). At this point the function is pertinent rather than the format of the belief.
- On the other side, as we have just argued, there is a co-referent Goal (wish, desire, intention, or any other motivational explicit representation).

Given the resulting amalgam these representations of the future are charged of value. Their intention or content has a (positive or negative) 'valence'[2]. More precisely, expectations can be:

**positive** (goal conformable):

$$Bel(x,p^{t'})^{t<t'} \& Goal(x,p^{t'}) \quad \left[or\ Bel(x,\sim p^{t'})^{t<t'} \& Goal(x,\sim p^{t'})\right]$$

**negative** (goal opposite):

$$Bel(x,p^{t'})^{t<t'} \& Goal(x,\sim p^{t'}) \quad \left[or\ Bel(x,\sim p^{t'})^{t<t'} \& Goal(x,p^{t'})\right]$$

**neutral**:

$$Bel(x,p^{t'})^{t<t'} \& \sim Goal(x,p^{t'}) \& \sim Goal(x,\sim p^{t'}) \quad \left[or\ Bel(x,\sim p^{t'})^{t<t'} \& \sim Goal(x,p^{t'}) \& \sim Goal(x,\sim p^{t'})\right]$$

**ambivalent**:

$$Bel(x,p^{t'})^{t<t'} \& Goal(x,p^{t'}) \& Goal(x,\sim p^{t'}) \quad \left[or\ Bel(x,\sim p^{t'})^{t<t'} \& Goal(x,p^{t'}) \& Goal(x,\sim p^{t'})\right]$$

## THE QUANTITATIVE ASPECTS OF MENTAL ATTITUDES

Decomposition of emotions in terms of beliefs and goals is not enough. We need quantitative' parameters. Frustration and pain have an intensity which can be more or less severe; the same holds for surprise, disappointment, relief, hope, joy etc. Since they are clearly related with what the agent believes, expects, likes, pursues, can we account for those dimensions on the basis of our (de)composition of those mental states, and of the basic epistemic and motivational representations? We claim so.

Given the two basic ingredients of any expectations (defined as different from simple forecast or prediction) Beliefs and Goals, we postulate that:

**P1:** Beliefs & Goals have specific quantitative dimensions; which are basically independent from each other.

Beliefs have strength, a degree of subjective certainty; the subject is more or less sure and committed about their content. Goals have a value, a subjective importance for the agent.

To simplify, we may have very important goals combined with uncertain predictions or pretty sure forecasts for not very relevant objectives etc. Thus, in our schematic notation, we should explicitly represent these dimensions of Goals and Beliefs:

$Bel^{\%}(x,p^t)$

$Goal^{\%}(x,p^t)$

where % represents the subjective importance or the value of the Goals and the subjective credibility and the certainty of the Beliefs.

Putting aside the Epistemic Goal, an expectation will be like this:

$Bel^{\%}(x,p^t)\&Goal^{\%}(x,\sim p^t)$

The subjective quality of those "configurations" or macro-attitudes will be very different precisely depending on those parameters. Also the effects of the invalidation of an expectation are very different depending on: (i) the positive or negative character of the expectation and (ii) the strengths of the components. Therefore, we postulate that:

**P2:** The dynamics and the degree of the emergent configuration or the macro-attitude are strictly a function of the dynamics and strength of its micro-components.

For example, when compared to the case of mere goal and high certainty, anxiety (Miceli and Castelfranchi, 2005) will probably be greater when the goal is very important and the uncertainty high and it is characterized by the need to know to reduce the uncertainty. In the following sections, we will characterize some of these emergent macro-attitudes.

### Hope and Fear

In our account, 'hope' is a peculiar kind of positive expectation where the goal is rather relevant for the subject while the expectation (more precisely the prediction) is not sure at all but rather weak and uncertain[3]:

$Bel^{low}(x,p^t)\&Goal^{high}(x,p^t)$

Correspondingly one might characterize 'fear', as an expectation of something bad, i.e. against our wishes:

$Bel^{\%}(x,p^t)\&Goal^{\%}(x,\sim p^t)$

But it seems that there can be 'fear' at any degree of certainty and of importance.[4]

Of course, these representations are seriously incomplete. We are ignoring their affective and felt component, which is definitely crucial. We do

not represent here the body, its states and signals to the control system; we are just providing their cognitive skeleton.

## THE IMPLICIT COUNTERPART OF EXPECTATIONS

Since we introduce a quantification of the degree of subjective certainty and reliability of Belief about the future (the forecast) we get a hidden, strange but nice consequence. There are other implicit opposite beliefs and thus implicit expectation. For implicit belief we mean here a belief that is not written or contained in any database (short term, working, or long term memory) but is only potentially known by the subject since it can be simply derived from actual beliefs. For example, my knowledge that Buenos Aires is the capital of Argentina is an explicit belief that I have in some memory and I just have to retrieve it. On the contrary, my knowledge that Buenos Aires is not the capital of Greece is not in any memory, but can just be derived (when needed) from what I explicitly know. Until it remains implicit, merely potential, and until it is not derived, it has no effect in my mind. For instance, I cannot perceive possible contradictions: my mind is only potentially contradictory if I believe that p, I believe that q, and p implies Not q, but I didn't derive that Not q.

Now, a belief that "70% it is the case that p", implies a belief that "30% it is the case that Not p"[5]. This has interesting consequences on expectations and related emotions. The Positive Expectation that p, entails an implicit (but sometime even explicit and compatible) Negative Expectation, and vice versa:

$$Bel^{\%}(x, p^t) \,\&\, Goal^{\%}(x, p^t) \Rightarrow Bel^{\%}(x, \sim p^t) \,\&\, Goal^{\%}(x, p^t)$$

This means that any hope implicitly contains some fear, and that any worry implicitly preserves some hope. But also means that when one gets a relief because a serious threat which was strongly expected has not happened and the world is conforming to her desires, she also gets (or can get) some exultance. It depends on her focus of attention and framing: is she focused on her worry and evanished threat, or on the unexpected achievement? Inversely, when one is satisfied for the actual expected realization of an important goal, she also can get some measure of relief while focusing on the previous implicit worry. It is not necessary to feel both the given emotion (i.e. fear) and the complementary one (i.e. hope) in a sort of oscillation or ambivalence and affective mixture. Only when the belief is explicitly represented and the attention is focused on it at least for a moment the corresponding emotion can be generated.

### Emotional Response to Expectation: The Strength of Disappointment

As we said, the effects of the invalidation of an expectation differ depending on: a) the positive or negative character of the expectation; b) the strengths of the components. Given the fact that X has previous expectations, how does this fact change her evaluation of and reaction to a given event?

### Invalidated Expectations

We call invalidated expectation an expectation that results to be wrong. For instance, X now (t'') believes that NOT p at time t' while she expects that p at time t':

*Invalidating*: $Bel(x,p^{t'})^{t<t'} \Longleftrightarrow Bel(x,\sim p^{t'})^{t''>t'}$

This crucial belief is the 'invalidating' belief. Relative to the goal component it represents a "frustration", "goal-failure". It is the frustrating belief: I desire, wish, want that p but I know that not p:

*Frustration*: $Goal(x,p^t)\&Bel(x,\sim p^{t'})$

Relative to the prediction belief, it represents the 'falsification', 'prediction-failure':

*Invalidation*: $Bel(x,p^{t'})^{t<t'} \& Bel(x,\sim p^{t'})^{t''>t'}$

$Bel(x,p^{t'})^{t<t'}$ represents the former illusion or delusion (X illusorily believed at time t that at t' p would be true).

This configuration provides also the cognitive basis and the components of "surprise": the more certain the prediction the more intense the surprise (Lorini and Castelfranchi, 2006; Machedo et alli, 2009). Given positive and negative expectations and the answer of the world, that is the frustrating or gratifying belief, we have either confirmation of the expectation or disappointment or relief.

## Disappointment

Relative to the whole mental state of positively expecting that p, the invalidating & frustrating belief produces 'disappointment' that is based on this basic configuration (plus the affective and cognitive reaction to it):

*Disappointment:*

$Goal^{\%}(x, p^{t'})^{t\&t'} \ \& \ Bel^{\%}(x, p^{t'})^{t} \ \& \ Bel^{\%}(x, \sim p^{t'})^{t}$

At time t, X believes that at t' (later, t' > t) p will be true; but now – at t' – she knows that Not p, while she continues to want that p. Disappointment contains goal-frustration and forecast failure, surprise. It entails a greater suffering than simple frustration (Miceli and Castelfranchi, 1997) for several reasons:

i. for the additional failure;
ii. for the fact that this impact also on the self-esteem as epistemic agent (Badura's (1990) predictability and related controllability) and is disorienting;
iii. for the fact that losses of a pre-existing fortune are worst than missed gains, and long expected and surely expected desired situation are so familiar and sure that we feel a sense of loss.

When the belief is stronger and well-grounded, the surprise gets more disorienting and restructuring and the consequences becomes stronger on our sense of predictability. When the goal becomes more important, the subject gets more frustrated.

In Disappointment these effects are combined:

*The surer the subject is about the outcome & the more important the outcome is for her, the more disappointed the subject will be.*

The degree of disappointment seems to be a function of both dimensions and components[6]. It seems to be felt as a unitary effect. Let's examine 4 situations in this regard, as and answer to the following question:

- *"How much are you disappointed?"*
  - *"I'm very disappointed: I was <u>sure</u> to succeed"*
  - *"I'm very disappointed: it was very <u>important</u> for me"*
  - *"Not at all: it was not <u>important</u> for me"*
  - *"Not at all: I have just tried; I was <u>expecting</u> a failure"*.

Obviously, worst disappointments are those with great value of the goal and high degree of certainty. However, the surprise component and the frustration component remain perceivable and a function of their specific variables.

## Relief

Relief is based on a negative expectation that results to be wrong. The prediction is invalidated but the goal is realized. There is no frustration but surprise. In a sense relief is the opposite of disappointment: the subject was down while expecting something bad, and now feel much better because this expectation is invalidated.

Relief:
$Goal(x, \sim p^{t'}) \,\&\, Bel(x, p^{t'}) \,\&\, Bel(x, \sim p^{t'})$ [7]

*The harder the expected harm and the more sure the expectation (i.e. the more serious the subjective threat) the more intense the relief. More precisely, the higher the worry, the threat, and the stronger the relief. The worry is already a function of the value of the harm and its certainty.*

Analogously, joy seems to be more intense depending on the value of the goal, but also on how unexpected it is. More specifically, for us 'joy' is not simply some form of happiness or some satisfaction for a goal achievement. It implies some excitation, in other words some significant arousal, which is precisely due to the fact that either the reward is higher than expected or the trust, the estimated probability, was not so high. In both cases, there is not only an achievement but also a positive surprise: something unexpected. For example, 'Exultance' seems a kind of joy, but due to a 'victory' against some perceived opposition, resistance, difficulty.

A more systematic analysis should distinguish between different kinds of surprise (based on different monitoring activities and on explicit vs. implicit beliefs), and different kinds of disappointment and relief due to the distinction between 'maintenance' situations and 'change/achievement' situations (Lorini and Castelfranchi, 2006).

More precisely (making constant the value of the Goal) the case of loss is usually worse than simple non-achievement. This is coherent with the theory of psychic suffering (Miceli and Castelfranchi, 1997) that claims that pain is greater when there is not only frustration but disappointment (that is a previous expectation), and when there is loss, not just missed gains, that is when the frustrated goal is a maintenance goal not an achievement goal. However, the presence of expectation makes this even more complicated.

## APPRAISAL AND THE COGNITIVE STRUCTURE OF COMPLEX EMOTIONS

People's appraisal of the meaning of a given state of affairs for their well-being is concordantly assumed to be a condition for their experiencing an emotion (Frijda and Swagerman, 1987; Ortony, 1987). Each emotion would involve a particular kind of appraisal, as well as a specific set of action tendencies and (perceived) physiological changes.

Cognitive models of emotion should then try

- to identify the specific cognitive processes implied by different emotions,
- by analyzing the structure of beliefs and goals typical of each of them.

Our analysis addresses such cognitive components, both directly and, so to say, indirectly, through the cognitive devices or strategies people can employ to elicit or to cope with that feeling (Miceli and Castelfranchi, 1998).

The general anatomy (sub-components) of a complex emotion is as follows:

**EMOTION** ***of x[before/towards y]for/about O***
BELIEFS
1. Bel about O
2. Bel about y
3. Bel (-/+ Evaluation O), …,. Bel (-/+ Evaluation O),
4. Bel (-/+ Expectations O), …, Bel (-/+ Expectations O)

MONITORED GOALS
5. Goal related to O or y → result: FRUSTRATION or REALIZATION

**ACTIVATED GOALS** (Action tendencies or *'impulses'*)
6. Goal in response to…

BODILY SENSATIONS
PLEASANT/UNPLEASANT FEELINGS
EMOTIONAL DISPLAY

## Shame

Let's provide an instantiation of a complex emotion using the above framework. Shame is a quite relevant social (but not necessarily moral) emotion that is due to the worry for the failure and frustration of our goal of having a good face (image), of being well evaluated by the others that observe and judge us (Castelfranchi and Poggi, 1990).

We feel ashamed about something (O) and before somebody (Y) whose opinion about us we care of.

**SHAME** x before y for/about O
BELIEFS

1. Bel x O where O = (Predicate of x) *"I did act" / "I have feature f"*
2. Bel x (Knows y O) *"they know/might know O"*
3. Bel x (negativeEvaluation y O) *"for them O is a fault, is bad, is negative"*
4. Bel x (negativeEvaluation y x) *"my image is defective; they do not like me"*
5. Bel x (negativeEvaluation x O) *"O is a bad thing"* SHARED VALUE
6. Bel x (negativeEvaluation x x) *"I'm defective"*

MONITORED GOALS

7. Goal x (positiveEvaluation y x) *being well evaluated; esteem, good image* → FRUSTRATION

ACTIVATED GOALS

8. Goal x (reducing exposure)

As for the first belief, notice that it is not strictly necessary. In other words while shared interiorized values are absolutely necessary, there can be disagreement about the evaluated fact. Although blushing, X might be innocent; she didn't do anything wrong. It is possible to blush and feel ashamed for the mere suspect. This is why blushing is not a confession at all.

As for the goal of shame (the goal of having positive evaluations, from Y), we have to notice that how the more X cares of Y's judgment and the worst Y's evaluation is, the greater X's suffering and shame intensity.

The 5th ingredient is very important, which is the personal negative evaluation of O by X herself, in other words; value sharing. Therefore, this statement implies that both self- and social esteem is harmed.

One cannot be ashamed of O in front of Y if:

- he does not (at least unconsciously) sincerely share some NegativeEvaluation of O (Shared/ Interiorized Values).
- he does not care at all of Y's opinion (Goal of PositiveEvaluation from Y; face/image; esteem).

What mainly matters in SHAME is:

FRUSTRATION → 5. Goal x (positiveEvaluation y x)
being well evaluated; esteem, good figure/face.

## Emotional Display of Shame

The Emotional display (posture, eyes, front, blushing) is very coherent with this complex mental state. The meaning of its non-verbal discourse (posture, eyes, front, blushing) is:

- *I care for your judgment; I care of being accepted in the group*
- *I recognize my fault, imperfection, flaw; I sincerely agree about its negativity;*
- *I sincerely share your values; I'm not an alien or a provocative; (consider that blushing cannot be simulated or inhibited);*
- *I do not oppose to you; I do submit to you;*
- *I'm suffering for my defect and your judgment; I'm sorry (I'm already paying for this)*
- *Be clement.* (Castelfranchi and Poggi, 1990)

In a similar manner, we present the complex anatomy of 'guilt' in the appendix.

## The "Intersubjectivity" of Social Emotions

Apart from possible mirroring, empathy, identification, etc., that imply some shared sensation and feeling, it is important to underline the shared and mutual mental ground of social emotions, also in their 'cognitive anatomies'. A shared mind is also crucial but left out from the discussion here.

Notice for example how Shame – in our anatomy – presupposes shared mental representations:

- The belief about O is 'shared', following X (and X believes so) (1 & 2);
- The negative evaluation of O is necessarily shared (and X believes so);
- Also the goal of X being well evaluated (that is, for Y the goal that: X be good, correspond to the cultural standards) is shared.

Moreover, beliefs and goals are not only shared, but they are meta-represented (Y's mind in X's mind, and X's mind in Y's mind (following X)) and mutual. We do not fully represent this, for sake of simplicity.

- X believes that Y believes that she shares the value (and this is actually true, especially after X's blushing signal);
- X believes that Y believes/knows that X believes (1) (2); etc.

They have (and know/feel to have) the same values, the same beliefs, the same goals.

And this is not based on complex reasoning and inferences about the other's mind, but mainly is due to their sharing a given culture, with its scripts, norms, values, conventions and behavioral rules, and to the emotional/behavioral signals and their automatic understanding.

## THE GESTALT NATURE OF COMPLEX MENTAL STATES

We are in favor of a componential analysis of emotions (and in general of mental states and processes, like 'expectation', or 'need', or 'trust', or 'argument', etc.) and this is what we refer as the cognitive anatomy. It allows a systematic explicit model of the relationships within that object and among objects. However, one should always also care about accounting for the 'unitary' character of those mental-behavioral phenomena. Although atomically decomposable, those complex mental states have their own emergent, specific, non-reducible properties and functions.

For example, a prediction is not *per se* an expectation, because it must be considered within a possible frame. It is a matter of the Gestalt nature of complex mental states. The side of a square is a linear segment; but: is any segment the side of a square? Not *per se*, only if considered, imagined, within that figure, as a component of a larger configuration that changes it meaning/role. Analogously: a belief about a future event is just part of an expectation, but it acquires a special color and function within the expectation Gestalt. Expectation is not simply the sum of a belief and a goal.

The fact that emotions are analyzable in parts and components (shared by other phenomena) does not necessarily deny their possible uniqueness and unitary/global nature. In our vocabulary they are Gestalts; there is an emergent, self-organizing form, which is not reducible to its parts and to their specific properties and functions. Decomposing a Gestalt is not reducing it to its components.

The new mental entity constitutes a Gestalt both phenomenally speaking, and functionally speaking: the whole has psychological and behavioral effects, properties, and functions that are new and specific; not just the results of the effects of its isolated parts. Moreover, within this whole the constituents change their nature, acquire a new color (or a role): they are - for example - no longer

just segments but have the function of sides of the emerging form of square. In this sense, even the elements that precede and cause the formation of such new complex object, remain there as its parts; since they are no longer exactly the same object.

There is a synergy among the constituents which is bi-directional: from micro to macro (the global form and effects), and from macro to micro: the role/function within the whole, and the new perception of/perspective on the part. Not necessarily all the components are there; but this gives rise to different related emotions, or to variations of the same emotion from more simple and primitive forms to richer ones (Miceli and Castelfranchi, 2009).

For example, in which sense an expectation - that compound configuration of beliefs and goals - is a new mental object, a unitary object? Because it, as a whole, acquires new properties, effects, and functions that are not properties of its parts. Like a molecule has properties that are not properties of its component atoms. For example, the expectation as such (not simply the prediction, or the goal) is involved in decision making. Expectations as such are needed for formulating an intention. Expectations as such produce 'hope' or 'fear', 'disappointment' or 'relief'.

The same holds for more complex mental states like 'shame' or 'guilt': they have a lot of common components, or components shared with other emotions, but they have their own specific subjective experience, and specific and global functions and reactions.

## RI-EMBODYING EMOTIONS

We believe that cognitive models have put aside for too long the problem of subjective experience, of feeling something. In our view quite obviously, to feel something is necessarily somatic; it presupposes having a body (including a brain), and receiving some perceptual signal from it. You cannot experience or feel anything without a body. However, current approaches claiming the role of the body, and feelings, emotions, drives, (and several biological mechanisms) tend to put this as a radical alternative to cognition, as incompatible with the traditional apparatus of cognitive science (beliefs, intentions, plans, decision, and so on).

To fully characterize several important mental states (kinds of belief or kinds of goal, like needs and desires (Castelfranchi, 2007)) it is necessary to model the bodily information; but on the other side – as we try to argue in this contribution - also traditional mental representations are necessary.

### 'Felt' Mental States

Notice, for example, that we cannot feel goal, intention, objective, plan, aim! Why? And why on the contrary can we feel needs and desires? (and a bit extensively hopes, expectations, trust).

Our trivial answer is: because they involve some perceptual component, while not all mental representations (goals, beliefs, etc.) involve significant perceptually active components, but are more abstract, or more disembodied representations. What we mean is that we cannot feel a goal per se but we can feel some perceptual component related to having a goal (like some uneasiness, or some perceptual representation of the expected results). While notions like needs or desires focus precisely on these aspects/components (Castelfranchi, 2007), other goal-notions are more abstract and do not explicitly concern these perceptual aspects.

**Desires** imply some pleasure, but not only the pleasure experienced at the moment of the achievement of the goal and satisfaction of the desire. Desires imply a pleasure at the very moment of desiring something as a mental activity. It is a virtual reality pleasure. A true desire implies the anticipatory representation of the goal state in a sensory-motor format (let's say an image) and the simulation of the desired situation. This implies some (partial) imagined sensation (for example the taste of a food; the joy of a sexual encounter).

What you feel is this sensation: an anticipated part of the sensation you will experience; an illusory gratification. To desire is this, and this is why you can feel a desire while you cannot feel an intention. The term intention does not focus on the perceptual anticipatory representation of the result and of its perceptual components.

So our claim would be: always when we can use the word[8] 'to feel' some somatic marker[9] or some self-perception is involved. Probably this is too strong, since the language extends the use of words and introduces metaphors; but it should be basically true.

It is important to understand that the problem is not only to go beyond a cognitive/functionalist analysis of emotions to integrate other aspects, but the problem is that any functional explanation is incomplete if ignores the subjective or felt facet of emotions. The real problem is precisely the function of the internal perception, of the feeling of the bodily peripheral reactions and of the central response. Since a reactive system can do the job of an emotional system, why do we need emotions? Why do we need a system that perceives its own reactions? What is the role of this self-perception in the adaptive process?

The classical AI position about emotions enounced by Simon (1967) explains their function in terms of operating system interrupts prompting one processing activity to be replaced by another of higher priority, i.e. in terms of a reactive goal-directed system in an unpredictable environment. As Sloman and Croucher (1981) observe, the need to cope with a changing and partly unpredictable world makes it very likely that any intelligent system with multiple motives and limited powers will have emotions. We believe that this view is basically correct but seriously incomplete. This function is necessary to explain emotions but is not sufficient at all. In fact, to deal with this kind of functionality a good reactive system able to focus attention or memory and to activate or inhibit goals and actions would be enough. Current models of affective computing simply model the emotional behavior and the cognitive-reactivity function. Consider for ex. Picard's nice description of fear in a robot:

*In its usual, nonemotional state, the robot peruses the planet, gathering data, analyzing it, and communicating its results back to earth. At one point, however, the robot senses that it has been physically damaged and changes to a new internal state, perhaps named 'fear'. In this new state it behaves differently, quickly reallocating its resources to drive its perceptual sensors and provide extra power to its motor system to let it move rapidly away from the source of danger. However, as long as the robot remains in a state of fear, it has insufficient resources to perform its data analysis (like human beings who can't concentrate on a task when they are in danger). The robot communication priorities, ceasing to be scientific, put out a call for help. (Picard, 1997).*

What is lacking in this characterization of fear? Just the most typical emotional aspect: feeling. Feeling is a broader notion: we can feel a lot of things that are not emotions (for example needs). However, feeling is a kernel component of emotion: if we cannot feel x, we should/could even doubt that x is an emotion (Ortony, 1987). This puts out a serious question: since we can account for emotional functioning without modeling feeling, since a reactive change of the internal state, cognitive processing, and behavior is enough, why is feeling such a crucial component of human (and animal) emotions? Is it a mere epiphenomenon lacking any causal function in the process? Or which is its function and its reason?

We believe that computational models of emotions should answer precisely this theoretical question. Let us simply mention what we believe to be the main functions of the feeling component in emotion, i.e. of the fact the robot should sense those changes of its internal state and of its behavior. We believe that the main functions of feeling in emotions are the following ones:

- felt emotional internal states work as drives (Canamero, 1997) to be satisfied, i.e. to go back to the equilibrium (homeostasis) through action; Mower (1960) postulates that in learning, the animal learns precisely what behavior serves to alleviate the emotion associated to a given stimulus;
- felt emotional internal states work as positive or negative internal reinforcements for learning (they will be associated to the episode and change the probability of the reproduction of the same behavior); [10]
- felt emotional internal states associated to and aroused by a given scenario constitute its immediate, unreasoned, non-declarative appraisal (to be distinguished from a cognitive evaluation - Castelfranchi, 2000; Miceli and Castelfranchi, 2000)

In sum, the cognitivist dominant paradigm cannot any longer neglect the necessity for modeling subjective experience and feeling. The relation with the body seems to be crucial: beliefs, goals, and other mental (declarative) ingredients are necessary but not sufficient. For example, one cannot account for the intentional aspect of feeling the need for something without beliefs about what is needed and about the origin of some sensation of pain or uneasiness. Also a better and convincing functionalist analysis of emotions requires precisely explaining the functional role of feeling. Cognitive appraisal, modification of attention and cognitive processes, reactive changes of goals priorities, are not adequate.

## Ri-Embodying 'Hope' and 'Fear'

In this paragraph we want to give some hints about the possible ri-embodiment of these mental states, claiming that their 'cognitive anatomy' is correct but insufficient. We will try to explain how the feeling aspect should be integrated with the epistemic and motivational ones.

Our claim is that

i. those mental configurations may produce a reaction of the body (a 'motion' **M**): a bodily response to that mental/representational content or interpretation of the events[11];
ii. this bodily response is entero-perceived by the agent, there is a signal **S** from the body: subjective 'sensations' about what is happening into the body or internal environment;
iii. these **S** and **M** are recognized (or attributed) as due to that event (or rather, interpreted event).

Only this provides the full experience and state of having 'worries about', or 'being afraid of', or 'feeling fear for' the eventuality that p will happen.

The molecule $Bel^{\%}(x, p^t)$ & $Goal^{\%}(x, \sim p^t)$ is not enough for fear: Where are the quiver, the tremble or the stress? Or the tremor that can characterize joy, the trepidation of hope?

Let's call **S** the sensation arriving from our body reacting to the prospected idea of a serious threat. Suppose that the reaction of the body, **M**, is a tremble, quivering, and that x perceives back this signal of the status of his body, **S**; and that he interprets this reaction as related and due to that (bad) mental prospect or better to its content (the negative event p).

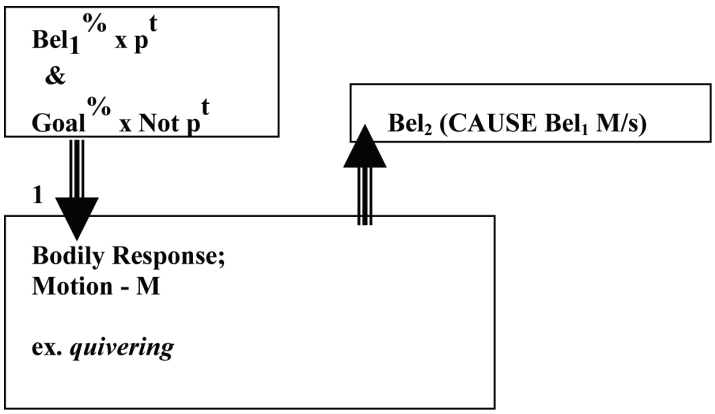
The complete picture is as follows:

Only at this point x really feels fear:

- on the one side, the simple negative expectation is affectively colored as 'fear', and
- on the other side, his tremor is a 'tremor of fear'.

Only the (causal) co-occurrence and association of the specific mental representation to the felt current bodily reaction (and possibly its cognitive attribution ($\mathbf{Bel_2}$)) accounts for what does it mean to worry about/for something, or to be afraid of something. Only the felt bodily reaction (feeling

*Figure 1. Beliefs - Body motion interaction*




and motion) makes this mental state emotion; but only the beliefs and goals provide the emotion with its origin.

The prediction is that: The greater the perceived threat- that is the more important the goal and the stronger the expectation- (moreover, the closer the check of the expectation, the verification, the expected moment **t**) the stronger the bodily reaction, the tremor or the tension.

## Real 'Hope'

Analogously, for having 'Hope' as a feeling, $Bel^{\%}(x, p^t)$ & $Goal^{\%}(x, p^t)$ is not enough. There is a reaction of the body (trepidation) to this prospect. And this trepidation is felt by X and related to that expectation. At that very moment X experiences hope (not just predicts a possible positive outcome). And only at that point (with this kind of body-mental-representation association) we have X's trepidation for hope.

The prediction is that: The more important the goal (and the closer the check of the expectation, the verification, the expected moment **t**) the stronger the trepidation.

## The Impact of the Felt Motion on Beliefs and Goals

But there are also other strange effects of the bodily feedback. The signal from the body can be used as an evidence, as a perceptual information source for the Belief. The intensity of the bodily sensation can affect the certainty of the forecast: '*Since I feel fear, there should be danger*'. Hence: The stronger the motion that x feels, the stronger the Belief.

However, this is an 'anomalous' and not very rational source. In fact its credibility has collapsed on its intensity. The degree of worries becomes, more broadly, a measure of the threat. A measure about not only the probability of the event (in our terms: degree of the prediction), but that is also the belief about the seriousness of the harm (i.e. the amount of the goal to be jeopardized).

The feeling might also affect the value of the goal: perhaps, the stronger the fear the greater the perceived value of the threatened Goal.

In general, as we saw, we claim that feeling provides an anomalous (nonrational) basis for both the strength of the Beliefs and values of the Goals. For example in felt 'needs for O' (Castelfranchi, 1998) we claim that: The stronger the disturbing or painful sensation that x feels when he feels 'the need for O', the stronger and more cogent and

*Figure 2. Bottom-up emotions*

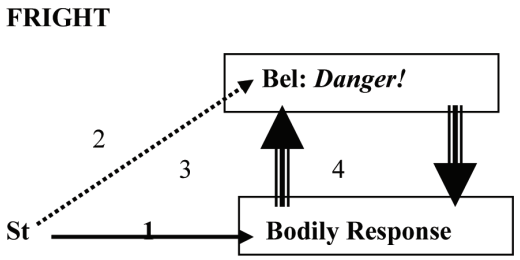
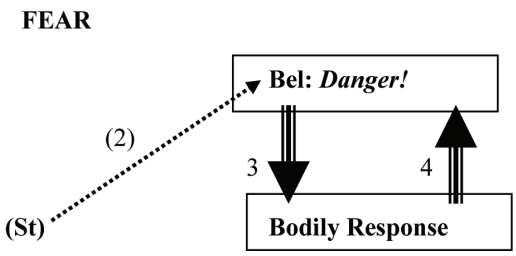



*Figure 3. Top-down emotions*




compulsory the Goal of 'having/obtaining O', that is not the usual way we calculate the importance of our Goals and their preference-order.

## Different Paths for Different Kinds of 'Fear'

We do not claim that this path (evaluation and mental interpretation of an event or a mental prospect m => bodily response **M** to m => sensorial feedback of the bodily response **S** => attribution of M & S to m) is the only path conducing to an emotional experience. There are more basic or primitive emotions that are more stimulus driven, not based on a match or mismatch between a Belief and a Goal (Reisenzein, 2009) like in shame which is presented before. A simple low level pattern matching is enough for eliciting an emotional reaction, for example a reactive fright to an unexpected noise (such as explosion) or to a non identified object suddenly moving/jumping under my feet. We claim that in this case there is no real evaluation and prediction of a possible danger. There is just an automatic (and sometimes conditioned) fear-reaction to the stimulus.

However, at least in humans, this motion of the body (and its sensations like being chilled with fear, or automatic retraction, horripilation, etc.) is a signal **S** that is interpreted and can generate a Belief of threat, and this Belief used as a feedback may confirm the bodily reaction. But the path is rather different. We have here some sort of Bottom-up (and back) emotions which is given in Figure 2 with the path 1 + 3 + 4 (In parallel, the cognitive processing of St proceeds on path 2).

While the previous flow was rather Top-down (and back), as shown in Figure 3.

Stimulus can even be absent; some fear can be just the result of a mere idea.

# EMOTIONS IN HUMAN MACHINE INTERACTION

## Emotional Cognition and Affective Interactions

We have provided earlier a quite brief anatomy of complex social emotions in order to give the reader the flavor of such specificity and complexity. However, actually even the anatomy of more simple affective attitudes like hope, fear, disappointment, relief, was rather cognitively rich.

In fact, we claim that the ascription of such a background mental state to the other is a necessary requisite for emotion recognition and understanding. In this section we will argue on this necessity of a Theory of Mind (ToM) for an appropriate affective response and interaction.

Appropriate emotional interactions are based on the recognition of the mental stuff of the other agent: of her beliefs, suppositions, motives, expectations. We react to this, not just to an expressive face, posture, or intonation. Expressive and physiological cues should mainly be the signs for a diagnosis of mind. Without this Theory of Mind map we are rather powerless.

Suppose that the other party is surprised; in order to appropriately react to this, I have to understand Why ? What for ? Moreover, is the other party just surprised or disappointed or relieved? What was her expectation, desire or worry? Is she ascribing to me the responsibility of such disappointment? I have to relate my response to this mental background. Should I express solidarity, excuse, or irony towards the other's expectation and reaction?

As we said, for an appropriate affective interaction, the mere detection and recognition of an emotional expression and a reaction to it is not enough. We do not react just to the emotion, but to the emotion as well as its 'aboutness'.

In other words, the affective reaction to an affective state depends on the recognition of the cognitive content (intension) of the affective state, not only on its expression (and the cultural and pragmatic context), which is summarized in Figure 4.

## Human Machine Interaction (HMI)

In sum, also for HMI: the detection and recognition of the symptom of an emotional state or reaction (speech prosody, heartbeat, facial expression) is not enough for an appropriate response. In fact, human and especially social emotions are based on a very rich and characteristic mental frame; on peculiar thoughts (the appraisal of current or possible events), on specific goals (frustrated or achieved), on action tendencies and activated impulses: what we call the "cognitive anatomy" of each emotion.

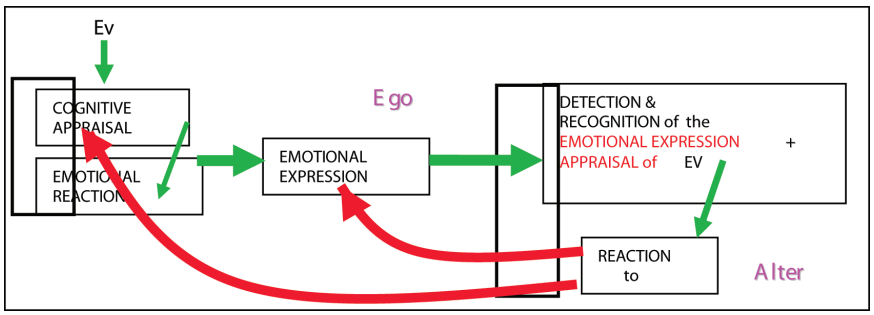
As for the HMI (and in particular H-Agent I) our claim is that:

- in HMI we are moving - also thanks to the Autonomous Agent (Ag) paradigm - from mere Interactivity to Collaboration;
- this more and more requires Ags able to have a some form of mind reading at least about the human user;
- thus, also the affective interaction cannot be merely behavioral and needs some level of mind-reading (if not felt empathy, shared sensations).

To model emotions, believable and appropriate faces and expressions are not enough. We have to build formal or computational models of the cognitive appraisal on which human emotions are based. The ability for Ags of building and reasoning upon explicit representations of the user's mind in terms of beliefs, expectations, desires, goals, needs, plans, values, etc.; in terms of social attitudes: like trust, diffidence, benevolence, hostility, etc.; and also in terms of the mental counterparts of speech acts, conventions, norms, roles, commitments, etc. is necessary for:

a. modeling credible emotional states as internal states;

*Figure 4. Reacting to the mental state not to the expressive signal*



b. modeling the sophisticated interaction between the cognitive components of the emotions (basically, beliefs and goals) and their bodily component: the felt 'motion' from the body; going beyond the quite schematic dual system models;
c. modeling a more credible affective interaction, where the detection and recognition of the symptoms of an emotional state or reaction (voice, heart, face,…) is not enough for an appropriate response to it, and the Ag must be able to understand the mental and subjective ground of the expressive reaction: what the emotion is about, that is what the user has in mind while feeling that emotion, what she believes, what she was desiring and expecting, what she would like or is pushed to do, etc.

Emotional interaction (Ag-Ag; H-Ag; H-robot; HC; etc.) cannot be based only on the recognition of the expressive or physiological signals.

Ideally we should have embodied the artificial minds; that is, modeling artificial agents with a real body, endowed with interoception and proprioception, bodily felt reactions, and an internal dynamic environment; and able to perceive and recognize the body response of the other and to react to it with a body felt response. Or even better: agents able to perceive the body response of the other through their own corresponding body activation. However, this would not be enough at all both for having emotional machines and for an affective HCI.

To make this idea concrete, let us give just an example that might be directly applied to interaction with anthropomorphic Agent: the appropriateness of an empathic response.

Even an emphatic response is not always the right, appropriate response. This strongly depends on the interpretation of the intention of the other's expressed emotion. For example, if X expresses irritation and rage against Z (a third part), then an empathic and sympathetic response of Y, sharing X's emotion, can be appropriate. This expresses support, solidarity, sympathy. But if the object of X's disappointment is precisely Y, Y's solidarity can be very inappropriate and irritating. The problem is: 'about what' and 'against who' X is furious? Is it about something that I personally did or provoked? Or is against X herself or a third agent?

If Y is the cause and the target of X's disappointment, a feeling of guilt, regret, and excuses could be much more appropriate than empathy in strict sense.

More precisely: suppose that X is disappointed and irritated against Y, this means that

Bel1 x (Done y act)
Bel2 x (Cause $act_y$ ev1)
Bel3 x (Harm ev1 x)
With a consequent negative evaluation of Y
Bel4 (negEVALUATION x y)

If Y shares those beliefs, if he agrees about Bel1, Bel2, Bel3 he can react by expressing guilt, sorrow ("*I'm sorry*") and excuses. But if Y disagrees about some of these Cognitive Appraisal ingredients, a different affective response is needed. For example, Y might be offended by X 'accusation': *"It is not my fault! How can you think this of me!"* (disagreement about at least Bel1 or Bel2, and thus Bel4). If Y *disagrees* about Bel3, the response can be completely different, not only verbally (*"But it is not bad! You didn't realize what really happened"*) but also as affective disposition: surprise and contrast. Suppose that X is disappointed and irritated against herself, the affective reaction of Y should again be quite different: *"It is not your fault" "it happens!"...* (friendly solidarity, consolation; or irony) And so on.

How can an Agent appropriately react to a perceived emotional state of the user without understanding what the user has in mind?

In sum: the affective reaction to an affective state depends on the recognition of the cognitive

content (intention) of that affective state, not only on its expression. We do not react to the emotion, but to the emotion and its 'aboutness', which presupposes some mind reading ability.

Moreover, we claim the emotional signals and expressions communicate also about this: about the mental content (beliefs, goals) not only about the affective feeling and disposition of the subject. Inferences from behaviors (implicit behavioral communication; Tummolini et al, 2004) cooperate with the specialized expressive signals to make us understand the emotional state of the subject, its reasons, and what it is about. For example, from your face I recognize that you are furious, but perhaps only from your behavior I realize that you are furious against me.

## CONCLUDING REMARKS (FOR HUMAN-AGENT INTERACTION)

The general conclusion is that we need a synthetic model of mental activity (and of emotion), able to assemble in a principled way both abstract and embodied representations, cognitive and dynamic dimensions. This is also necessary for the theory of emotions and for the theory of motivation that cannot be reduced to their bodily components, arousal, impulses, etc. but require specific beliefs, goals, expectations, explicit evaluations, and so on.

This strongly impacts the affective interaction too, which cannot be reduced to (and just modeled as) a felt empathic reaction: disgust elicits disgust, suffering elicits suffering, and so on; but requires some explicit mind reading and some appropriated reaction to the other's mental assumptions: beliefs, expectations, goals .

Do we really want a social interaction with our artificial creatures? Do we really want to support and mediate human interaction by the computer technology? There is no alternative: we have to explicitly and computationally model the mental proximate mechanisms generating the behaviors (both, the affective and the more reasoned and deliberated ones), and we have to address our response or our support-mediation to them, not just to the exterior behaviors and signals. No emotions without cognition and motivation, no interaction without understanding.

## ACKNOWLEDGMENT

I would like to thanks Maria Miceli, Isabella Poggi, Rino Falcone, Emiliano Lorini, Luca Tummolini, Giovanni Pezzulo for their invaluable contribution to many of the ideas presented here.

## REFERENCES

Bandura, A. (1990). Self-efficacy mechanism in human agency. *The American Psychologist*, *37*, 122–147. doi:10.1037/0003-066X.37.2.122

Canamero, D. (1997) Modeling Motivations and Emotions as a Basis for Intelligent Behavior. *Autonomous Agents'98*, ACM Press, 148-55, 1997.

Castelfranchi, C. (1998). To believe and to feel: The case of "needs". In Canamero, D. (Ed.), *Emotional and Intelligent: The Tangled Knot of Cognition* (pp. 55–60). Menlo Park, CA: AAAI Press.

Castelfranchi, C. (2000). Affective Appraisal vs. Cognitive Evaluation in Social Emotions and Interactions. In A. Paiva (ed.) *Affective Interactions. Towards a New Generation of Computer Interfaces*. Heidelbergh, Springer, LNAI 1814, 76-106.

Castelfranchi, C. (2005). Mind as an Anticipatory Device: For a Theory of Expectations. BVAI 2005 Brain, vision, and artificial intelligence (First international symposium, BVAI 2005, Naples, Italy, October 19-21, 2005) (proceedings). *Lecture Notes in Computer Science*, *3704*, 258–276. doi:10.1007/11565123_26

Castelfranchi, C., & Miceli, M. (2009). The Cognitive-Motivational Compound of Emotional Experience. *Emotion Review*, *1*(3), 221–228. doi:10.1177/1754073909103590

Castelfranchi, C., & Poggi, I. (1990). Blushing as a discourse: was Darwin wrong? In Crozier, R. (Ed.), *Shyness and Embarassement: Perspective from Social Psychology*. N. Y: Cambridge University Press. doi:10.1017/CBO9780511571183.009

Damasio, A.R. (1994) *Descartes' Error*. N.Y., Putnam's Sons, Frijda N. H. & Swagerman J. (1987) Can Computers Feel? Theory and Design of an Emotional System. *Cognition and Emotion, 1* (3) 235-57, 1987.

Lorini, E., & Castelfranchi, C. (2006). The Unexpected Aspects of Surprise. *IJPRAI*, *20*(6), 817–834.

Macedo, L., Cardoso, A., Reisenzein, R., Lorini, E., & Castelfranchi, C. (2009). Artificial Surprise. In Vallverdú, J., & Casacuberta, D. (Eds.), *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*.

Miceli, M., & Castelfranchi, C. (1997). Basic principles of psychic suffering: A prelimirary account. *Theory & Psychology*, *7*, 769–798. doi:10.1177/0959354397076003

Miceli, M., & Castelfranchi, C. (1998). How to Silence One's Conscience: Cognitive Defences Against the Feeling of Guilt. *Journal for the Theory of Social Behaviour*, *28*(3), 287–318. doi:10.1111/1468-5914.00076

Miceli, M., & Castelfranchi, C. (2000). The role of evaluation in cognition and social interaction. In Dautenhahn, K. (Ed.), *Human cognition and agent technology* (pp. 225–261). Amsterdam: Benjamins.

Miceli, M. e Castelfranchi, C. (2002). The mind and the future: The (negative) power of expectations. *Theory & Psychology,*12 (3), 335-.366

Miceli, M., Castelfranchi, C. (2005) Anxiety as an epistemic emotion: An uncertainty theory of anxiety, *Anxiety Stress and Coping* (09957J0), 18,291-319.

Mower, O. (1960). *Learning Theory and Behavior*. New York: J. Wiley and Sons. doi:10.1037/10802-000

Ortony, A. (1987) Is Guilt an Emotion? *Cognition and Emotion*, I, 1, 283-98, 1987.

Pezzulo, G., Butz, M. V., Castelfranchi, C., & Falcone, R. (2008). Anticipation in Natural and Artificial Cognition. In Pezzulo, G., Butz, M.V., Castelfranchi, C. & Falcone, R. (Eds.), *The Challenge of Anticipation: A Unifying Framework for the Analysis and Design of Artificial Cognitive Systems* (LNAI 5225, pp. 3-22). New York: Springer.

Picard, R. (1997). *Does HAL cry digital tears? Emotion and computers;* HAL's Legacy: 2001's Computer as Dream and Reality, Cambridge, 279-303.

Reisenzein, R. (2009). Emotional Experience in the Computational Belief–Desire Theory of Emotion. *Emotion Review*, *1*(3), 214–222. doi:10.1177/1754073909103589

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review*, *74*, 29–39. doi:10.1037/h0024127

Sloman, A., & Croucher, M. (1981) Why robots will have emotions. In Proceedings of *IJCAI'81*, Vancouver, Canada, 1981, p. 197

Staats, A. (1990). The Paradigmatic Behaviorism Theory of Emotions: Basis for Unification. *Clinical Psychology Review*, *10*, 539–566. doi:10.1016/0272-7358(90)90096-S

Tummolini, L., Castelfranchi, C., Ricci, A., Viroli, M., & Omicini, A. (2004) What I See is What You Say: Coordination in a Shared Environment with Behavioral Implicit Communication. In G. Vouros (Ed.) *ECAI 04 Proceedings of the Workshop on Coordination in Emergent Societies* (CEAS 2004).

## ENDNOTES

1 'Prediction' is the result of the action of predicting; but 'expectation' is not the result of the action of expecting; it is that action or the outcome of a prediction relevant to goals, it is the basis of such an action.

2 Actually the expectation entails a cognitive evaluation. In fact, since the realization of p is coinciding with a goal, it is 'good'. If belief is the opposite of the goal, it implies a belief that the outcome of the world will be 'bad'. Or the expectation produces an implicit, intuitive appraisal, simply by activating associated affective responses or somatic markers (Damasio, 1994). Or the expected result will produce a reward for the agent, and – although not strictly driving its behavior- it is positive since it will satisfy a drive and reinforce the behavior.

3 We may also have 'strong hope' but we explicitly call it strong precisely because usually hope implies low confidence and some anxiety and worry. In any case, 'hope' (like explicit 'trust') can never really be subjectively certain and absolutely confident. Hope implies uncertainty. More precisely, hope should be based on a belief of possibility rather than on an estimated probability of the event.

4 To characterize *fear* another component would be very relevant: the goal of avoiding the foreseen danger; that is, the goal of *doing* something such that Not p. This is a goal activated while feeling fear. But it is also a component of a complete fear mental state, not just a follower or a consequence of fear. This goal can be a quite specified action (motor reaction) (a cry; the impulse to escape; etc.); or a generic goal 'doing something' ("my God!! What can I do?!"). The more intense the felt fear, the more important the activate goal of avoidance.

5 In fact it is possible that there is an interval of ignorance, some lack of evidences; that is that I estimate with a probability of 45% that p and with a probability of 30% Not p, while having a gap of 25% neither in favor of p nor of Not p.

6 As a first approximation of the degree of Disappointment one might assume some sort of multiplication of the two factors: Goal-value * Belief-certainty. Similarly to 'Subjective Expected Utility': the greater the SEU the more intense the Disappointment.

7 Or – obviously - (Goal x pt') & (Bel x ¬pt') & (Bel x pt').

8 This is especially true in Italian (the semantic difference between "sentire" and "provare"); perhaps less true in English where really to "feel" seems quite close to "believe".

9 In Damasio's terminology (Damasion, 1994) a somatic marker is a positive or negative emotional reaction in the brain that is associated to and elicited by a given mental representation or scenario, making it attractive or repulsive, and pre-orienting choice. It may just be the central trace of an original peripheral, physiological reaction.

10 I assume, following along tradition on emotional learning, that in general positive and negative emotions are reinforcers; but notice that this does neither imply that we act in order to feel the emotion, which is not necessarily motivating us (it can be expected without being intended); nor that

only pleasure and pain, or emotions, are rewarding (Staats, 1990).

11 As we said, *also this* makes them like 'molecules' with their own global properties and effects, since the 'response' is to the whole pattern not to its components, and it is not just the sum of the specific reactions to the components.

## APPENDIX

### Anatomy of Guilt

The prototypical kind of Guilt is an unpleasant feeling, a sufferance for having been responsible of some harm to a victim. One feels guilt about something (O) and for somebody (Z) who is suffering or might suffer for that harm (Miceli and Castelfranchi, 1998).

**GUILT** x for/about O harming z [before y/z]
BELIEFS

1. Bel x O were O = (Did x act) *"I did act1"*
2. Bel x (Cause O (Fate of z)) *"act1 affected y's fate/condition"*
3. Bel x (Harm for z) *"it is a bad fate, a harm for z"*
4. Bel x (will/could Suffer z) *"z is suffering; will suffer; might suffer"*
5. Bel x (not deserved by z) *"z is a victim, did not deserve this harm"*
6. Bel x (Could have avoided x act) *"I could have avoided this"* (counterfactual)
7. Bel x (-Evaluation y O) *"for them O is a wrong, is bad, is negative"*
8. Bel x (-Evaluation y x) *"I'm a bad guy for him/them"*

MONITORED GOALS

9. Goal x: (Not being cause of an unfair harms)
10. Goal x (+Evaluation y x) *being well evaluated; moral estime, moral image*
11. -Evaluation x O *"O is a bad thing"* SHARED VALUE
12. -Evaluation x x "*I'm not good; I'm a bad guy"*

**ACTIVATED GOALS** (Action tendencies or *'impulses'*)

13. Goal (help x y*) "compensating; worrying about; to care of y*" →ANXIETY
14. Goal x (Expiate x) *"to atone; pay for…"* → ANXIETY
15. Goal x (Not Did x act*) counterfactual desire*; REGRET (IMPOSSIBLE Goal!)
16. Goal x (Not Does X act in the future): virtuous intention

Let's remark that guilt feeling presupposes the capacity for empathy: belief (4.) and (5.) activate an empathic attitude; X imagines and feels Y's sufferance, and this is one of the basis of Guilt intensity (the other are the degree of responsibility, the perceived gravity of the harm, the degree of unfairness).

Very crucial is also the counterfactual belief (6.); it is the core of the sense of responsibility. Moreover: since I could have avoided my act or the harm, I should have avoided it! (This is the internal reproach, the remorse, and also the basis for the good intention for the future).

To be more precise (6.) is a group of related beliefs: like "*I could/should have understood the consequences*"; "*I had some freedom; I was not forced to do so*" .

Also guilt feeling implies a negative evaluation of the action and of X (and thus a wound to self-esteem and – if somebody can know and judge (but it is not necessary for Guilt) – also shared values and a wound to social image. However, this is not the goal guilt feelings monitor and are about.

Guilt mainly is about causal links between our own action or fate with the other's bad fate, and focuses on our bad power (the power to harm, to be noxious); while shame focuses on our lack of power, inferiority, inadequacy, and defectiveness; and on face problems. Shame elicits a passive and depressive

attitude; while Guilt an active and reparative attitude (M. Lewis, 1992). In guilt the preeminent role is played by the assumed responsibility.

One generally does not feel responsible for one's lack of power and inadequacy, which are often perceived as beyond one's control. Hence, the passive and depressive attitude. By contrast, one's injurious behavior, negative power and dispositions are seen as controllable and modifiable.

People can feel ashamed because of their ugliness or handicaps, but they don't feel guilty (unless they attribute themselves some responsibility for not trying enough to improve themselves or avoid bad consequences). Conversely, people tend to feel guilty, rather than ashamed, for their bad behavior or dispositions.

What mainly matters in GUILT is:

FRUSTRATION → 7. Goal: (Not being cause of an unfair harm)
moral self-esteem and reproach