Mind as an Anticipatory Device: For a Theory of Expectations

Cristiano Castelfranchi

Institute for Cognitive Sciences and Technologies - CNR
Via San Martino della Battaglia 44, 00185, Roma, Italy

Abstract

This work is about the central role of "expectations" in mental life and in purposive action. We will present a cognitive anatomy of expectations, their reduction in terms of more elementary ingredients: beliefs and goals. Moreover, those ingredients will be considered in their 'quantitative' dimension: the value of the Goal, the strength of the Beliefs. We will base several predictions on this analytical distinction, and sketch a theory of hope, fear, frustration, disappointment, and relief, strictly derived from the analysis of expectations. Eventually, we will discuss how we can capture the global subjective character of such mental states that we have decomposed; how to account for their gestaltic nature.

'More geometrico demonstrata'
Spinoza

Premise: The Anticipatory Nature of Mind

Basically mind is for "anticipation" (Butz & Hoffman 2002), or – more precisely – for building and working upon "anticipatory representations" (Miceli & Castelfranchi 2002; Castelfranchi & Lorini 2003). A real "mental" activity and representation starts to be there when the organism is able to endogenously (not as the output of current perceptual stimuli) produce an internal perceptual representation of the world (simulation of perception). Which is the origin and the use of such strange ability? There are several uses or functions but many (if not all) of them are anticipatory. For example, the organism can generate the internal "image" for matching it against perceptual inputs while actively searching for a given object or stimulus while exploring an environment; or can use it as prediction of the stimulus that will probably arrive, as in active 'recognition'. It can use the perceptual expectation for implicitly monitoring the 'success' of the rule-based, reactive behavior, and as criteria for reinforcing or not the rule (Butz & Hoffman 2002). But it can also entertain a mental representation of the current world just for working on it, modifying this representation for virtually 'exploring' possible actions, events, results: "what will/would happen if…?".

This precisely is "intelligence": not just the capacity to exhibit complex adaptive behaviors (like in social insects or in spiders), nor the capacity to solve problems (for example by stupid and blind trial and errors!), but the capacity to solve a problem by working on an internal representation of the problem, by acting upon 'images' with simulated actions, or on 'mental models' or 'symbolic representations' by mental actions, transformations ('reasoning'), before performing the actions in the world. The architect designs in her mind (and on a piece of paper) her building before building it; this is not the case of a spider although what it will build will be very complex (and - for us - beautiful).

Those mental representations that characterize the mind and the mental work are mainly for anticipation: before the stimulus to be matched (prediction), before the action to be executed (project), etc. This means that the ability that characterizes and defines a "mind" is that of building representations of the non-existent, of what is not currently (yet) "true", perceivable.

This clearly builds upon memory, that is the re-evocable traces of previously perceived scenes; usually is just past "experience" evoked and projected on the future. But this is only the origin. A fully developed mind is able to build never-seen scenes, new possible combinations of world elements never perceived; it is a real building and creation (by simulation) not just memory retrieval.

Moreover, the use of such internally and autonomously generated representations of the world is not only "epistemic", for knowledge of the past, the present, the future: that is memory, perception, prediction and expectations. Those representations can have a radically different function: they can have motivational, axiological, or deontic nature; saying us not how the world is, was, will be; but how the world should be, how the organism would like the world to be. That is these representations can be used as goals driving the behavior. While an adaptive organism tends to adjust its epistemic representations (knowledge; beliefs) to the "reality", to make their fidelity to the world as much accurate as possible; on the opposite an effective goal-directed system try to adjust the "objective" external world to its endogenous representation! To change the world (through the "action" which in fact is goal-directed behavior) and make it the closest possible to its internally creative mental picture (that could be a picture of something never existed)! This really is a "mind": the presupposition for hallucinations, delirium, desires, and utopias.

Like "signs" are really signs when they can be used for deception and lie, not when they just are the non-autonomous index of reality, propagating from it; analogously, mental representations (that in fact - as any "representation" - are complex "signs") are really there were they can be false and independently generated from reality. The use of this anticipation is not only for prediction (by definition the future is currently not-true) but also, more importantly, for the purposive character of the behavior, for internal explicit goal representation[1].

Steps in Anticipation: Anticipatory Behaviors vs. Anticipatory Representations

Any purposive behavior (in strict sense), any goal-directed system is necessarily anticipatory, since it is driven by the representation of the goal-state (set-point) and activated by its mismatch with the current state of the world (Rosenblueth *et al.* 1960; Miller *et al.* 1960). But not any anticipatory behavior, and even not necessarily any behavior based on anticipatory representations, is goal-directed.

As for the claim relative to the fact that not any anticipatory behavior (Butz 2002) is based on explicit cognitive representations of future relevant/concerning events, that is on expectations, one should consider many instances of 'implicit' or merely behavioral anticipation or preparation, where the agent simply 'reacts' to a stimulus with a behavioral response (conditioned/learned or unconditioned/inborn) but in fact the response is functional, apt to some incoming event. The stimulus is some sort of 'precursory sign' and the response in fact is preparatory to the 'announced' event:

Precursory stimulus ⇒ Preparatory behavior ⇒ Event
(e.g. noise) (e.g. jump) (e.g. approaching predator)

In this case there is no explicit 'mental' representation of the future event. It is just a case of what we propose to call 'merely anticipatory behavior'. A Stimulus St is exploited (thanks to selection or learning) as the precursor and the 'sign' of a following event Ev, and it is adaptive for the organism to respond immediately to St with a behavior which in fact is just the 'preparation' to the forthcoming Ev; the advantage is that the organism is 'ready', 'prepared to' Ev. But this does not require a 'mental' anticipated explicit representation of Ev, that is the prediction, or better the 'expectation' that Ev will occur.

Surprise. The first level of cognitive anticipation is the retrieval from memory of previous perceptual experience to be compared with the incoming perceptual input (some sort of procedural 'prediction'). The use of this perceptual anticipation is multiple.

On the one side it is applied not only to action but also to the processes of the world and it is for monitoring the course of the events. Its function seems to be that of detecting unusual events that might require additional epistemic processing (for example attention) or a fast reaction. One might claim that even before this, clearly any form of pattern matching (where the pattern is either inborn or learned) is an *implicit* form of anticipation since it should be based on past experience and -more importantly- should fit some features of the environment, should be adapted to it, thus implicitly expecting and predicting given features in the environment (Bickhard 2005). Beyond this, there are true predictions activated by premonitory signs that 'announce' a given event.

The function of this systematic monitoring of the world is also of continuously updating and readjusting the world representation, to see whether predictions are correct and pertinent and the world can be 'assimilated' to current schemata or if it is the case to have some 'accommodation' of them (Piaget).

On the other side, the internal simulation of the next percept is fundamental for teleonomic behavior; during the action it is crucial to compare the perceptual feedback (both proprioceptive and external) with some representation of the expected state of the body and of the world. Indeed, we can argue that whenever an agent executes an action there is at least an automatic not intentional perceptive test on the success of the action. This idea is supported from empirical research and is a building block in neuro-psychology inspired computational models of action control (see Jordan & Wolpert 1999 for a review): the importance of sensory feedback for the adjustment of the goal-directed motor behaviour in phase of action execution. Only this match or mismatch (after the test) can say to the agent if there is something wrong (Ortony & Partridge 1987).

This kinds of sensory-motor expectations already allows some form of 'surprise', the most peripheral one, just due to perceptual mismatch; a first-hand surprise. **'Surprise'** is the automatic reaction to a mismatch. It is:

- a (felt) reaction/response

- of alert and arousal

- due to an inconsistency (mismatch, non-assimilation, lack of integration) between incoming information and our previous knowledge, in particular an actual prediction or a potential prediction

- invoking and mobilizing resources at disposal of an activity for a better epistemic processing of this 'strange' information (attention, search, belief revision, etc.),

- aimed at solving the inconsistency

- and at preventing possible dangers (the reason for the alarm) due to a lack of predictability and to a wrong anticipation.

[1] In this perspective the "omeostatic" view of goals and of their cybernetic, feed-back machinery is a bit misleading. "Omeo-stasys" gives the idea of maintaining and restoring an existing state that can be disturbed; but in fact the cybernetic model and the notion of goal refer also to the instauration of states that have never been there! This is why the notion of "purposive" behavior is much better, although definitely founded on the same model.

The deeper and slower forms of surprise are due to symbolic representations of expected events, and to the process of information integration with previous long-term knowledge. This is surprise due to *implausibility*, unbelievability of the new information (Lorini & Castelfranchi in preparation).

In this work we mainly focus on true predictions (based on inference, reasoning, mental models) (although they can also be mental 'images' in sensory format), and on their combination with explicit goals to produce the specific mental object called 'Expectation'.

Low level 'predictions' are based on some form of 'statistical' learning, on frequency and regular sequences, on judgment of normality in direct perceptual experience, on the strength of associative links and on the probability of activation (Kahneman & Miller 1986).

High level predictions have many different sources: from analogy ("The first time he was very elegant, I think that he will be well dressed") and, in general, inferences and reasoning ("He is Italian thus he will love pasta"), to natural laws, and – in social domain - to norms, roles, conventions, habits, scripts ("He will not do so; here it is prohibited"), or to "Theory of Mind" ("He hate John, so he will try to…"; "He decided to go in vacation, so he will not be here on Monday").

Proto-expectations. As for anticipatory-representation-based behaviors that are not strictly goal-directed (intention like) let us briefly discuss also a weaker and more primitive form of 'expectation'; the anticipatory representation of the result of the action in 'Anticipatory Classifiers' (AC) (Butz 2002; Drescher 1991). In our interpretation, they are not simply 'predictions'. They represent a forerunner of true Expectations because the agent is not unconcerned, but it actively checks whether the prediction is true, because the result is highly relevant, since it satisfies (or non-satisfies) a drive, and provides a reward. But on the other side, for us – in their basic form- they can (and should) be distinguished from true 'goal' in the classical 'purposive behavior' sense (Rosenblueth *et al.* 1960)

As we just said Expectations should be distinguished from various forms of mere anticipation and of behavioral preparation. These are the implicit and procedural forerunners of true cognitive expectations. These are pseudo-expectations: the agent behaves "as if" it had an expectation. Consider for example unconditioned salivation in Pavlov experiments. This is just a preparatory reaction for eating. It is based on a current stimulus eliciting a response that is useful (a condition) for a future behavior: preparation. Consider automatic coordination (either inborn or learned) in swallowing or walking, or in dodging a flying rock. Finally, consider our implicit and procedural trust that the ground will not sink under our feet, or that water is liquid, and snow cold, etc. In some case there is no representation at all; but simply a default behavior or procedure: the expectation is the lack of special control (i.e. of the ground).

However, in other cases there is the *anticipatory representation* internally generated, simulated, of a sensation (perceptual input) which will be compared with the actual one. This is very close to an Expectation (at least to its Prediction component); however, there is no necessarily an explicit real Goal initiating the process, searching for the action, and a purposive-behavior feedback, for monitoring and adjusting the action. A simple AC is enough. An AC can just remain a production rule, a classifier, something close to a stimulus-response link, that has also (in the right part) some representation of the predicted/learned result.

Cond ==> Act + ExpResult

This representation is compared against the actual result: if it matches (correct expectation) the links (between Cond and Act and between Act and ExpResult) will be reinforced; if it does not match (wrong prediction) the rule will be weakened.

We assume that this (which for us too is the device underlying Skinner's 'instrumental learning'; Butz & Hoffman 2002) in not necessarily yet 'purposive behavior' and that the expected result (ExpResult) is not really a Goal (like in the TOTE model). The behavior is data/input driven, rule-based, not explicitly 'purposive', not top-down elicited and guided by the representation of its Goal, and cannot be creative and new, cannot start a problem-solving activity.

In this paper we will model only explicit anticipatory representations, and in particular Expectations in strong sense, and their role in a goal-directed mind and intentional behavior. We will present a *Cognitive Anatomy* of Expectations, their reduction in terms of more elementary ingredients: beliefs and goals; and their 'strength'. We will base several predictions on this analytical distinction. We will present a theory of hope, worries, frustration, disappointment, relief, ready for artificial creature: could robots and software agents move from low level form of anticipation, surprise, etc. to explicit expectations and related mental states?

Let us start by disentangling simple predictions from true expectations.

Cognitive Anatomy of Expectations

Prediction vs. Expectation

'Expectation' is not synonymous of 'prediction' or 'forecast'; they have a common semantic core (a belief – more or less certain[2]– about the future[3]) and thus a

[2] In some Dictionary 'Expectation' is defined as: "1. a *confident* belief or *strong hope* that a particular event will happen" (Encarta® World English Dictionary © 1999 Microsoft Corporation). Notice also the positive connotation of the expected event (hope), while in fact also

partially overlapping extension. We consider a forecast (Miceli & Castelfranchi 2002; Castelfranchi & Lorini 2003) as a mere belief about a future state of the world and we distinguish it from a simple ‘hypothesis’. The difference is in term of degree of certainty: a hypothesis may involve the belief that future p is possible while in a forecast the belief that future p is probable. A forecast implies that the chance threshold has been exceeded (domain of probability). According to the agent’s past experience or knowledge of physical or social rules and laws p should happen (in an epistemic sense)[4].

Putting aside the degree of confidence (we need a general term covering weak and strong predictions), one might say that EXPECTATION ➔ PREDICTION, or better that both of them imply a representation of a possible future: a possible Belief about the future. But they also have different features. The primary difference is that in ‘expectation’ (but not necessarily and conceptually in ‘prediction’) there is also a motivational component; some Goal of the subject X is involved. X is ‘concerned’: she didn’t just ‘predict’ and be indifferent to the event or mindless. Let’s carefully analyze this motivational and active component.

Epistemic Goals and Activity. First of all, X has the Goal to know 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 of ‘looking for’ is necessarily related to the notion of expecting and expectation, but not to the notion of prediction.

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 in any case at the moment of the future event or state compare what happens with her prediction (epistemic actions) (Lorini & Castelfranchi 2004; Kirsh & Maglio 1994). Because in any case she has the Goal to know whether the world actually is as anticipated, and if the prediction was correct. Schematically [5]:

Expectation x p ➔
Bel x at t’ that p at t” (where t” > t’)
Goal x from t’ to t’” KnowWhether x p or Not p at t”
(t’” ≥ t”)

This really is ‘expecting’ and the true ‘expectation’.

Content Goals. This Epistemic/monitoring Goal is combined with Goals about p: the agent’s need, desire, or ‘intention that’ the world should realize. The Goal that p is true (that is the Goal that p) 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 [6].

When the agent has a goal opposite to her prediction, she has a ‘negative expectation’; when the agent has a goal equal to her prediction she has a ‘positive expectation’. To be true a Goal equal to the prediction in Expectation is always there, although frequently quite weak and secondary relatively to the main concern. In fact, when X predicts that p and monitors the world to know whether actually p, she has also the Goal that p, just in order to not disconfirm her prediction, and to confirm to be a good predictor, to feel that the world is predictable and have a sense of ‘control’. We are referring to *predictability*, that is, the cognitive component of self-efficacy (Bandura 1990): the need to anticipate future events and the consequent need to find such anticipation validated by facts. This need for prediction is functional in humans in order to avoid anxiety, disorientation and distress. Cooper and Fazio (1984) have experimentally proved that people act in order to find their forecasts (predictions) validated by facts and feel distressed by invalidation.

Defining Expectations

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 the 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 the pre-vision (to for-see).

‘negative or bad’ expectations are possible (worries). Notice also the second definition: “2. a mental image of something expected, often compared to its reality” where both the nature of an explicit mental representation, and the monitoring/epistemic activity are correctly identified.

[3] Also predictions and expectations about the past are possible but only in the sense that one will come in the future to know something about the past and has some hypothesis and wish on that.

[4] Consider for example the definition of ‘forecasting’: “to predict or work out something that is likely to happen…” (Encarta® World English Dictionary © 1999 Microsoft Corporation.)

[5] We will not use here a logical formalization; we will just use a self-explanatory and synthetic notation, useful for a schematic characterization of different combinations of beliefs and goals. For a real formalization of some of these mental attitudes see Castelfranchi & Lorini 2003.

[6] For example, computers make weather ‘forecasts’ but it would be strange to say that they ‘have expectations’ about the weather. Currently they are ‘unconcerned’.

The format of this belief or pre-vision can be either propositional or imagery (or mental model of); this does not matter. Here just the function is pertinent.

- On the other side, as we 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 'valence': it is positive, or negative, and so on.

- Either, the expectation entails a cognitive evaluation (Miceli & Castelfranchi 2000).

In fact, since the realization of p is coinciding with a goal, it is "good"; while if the 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 (Miceli & Castelfranchi 2000); or both;
- Or the expected result will produce a *reward* for the agent, and – although not strictly driving its behavior, it is positive for it since it will satisfy a drive and reinforce the behavior.[7]

We analyze here only the Expectations in a strong sense, with an explicit Goal; but we mentioned Expectations in those forms of reactive, rule-based behaviors, first in order to stress how the notion of Expectation always involves the idea of a *valence* and of the agent being concerned and monitoring the world; second, to give an idea of more elementary and forerunner forms of this construct.

Positive and Negative Expectations

Expectation can be:

- **positive** (goal conformable): $(\text{Bel } x\ p^{t'})^{t<t'}$ & $(\text{Goal } x\ p^{t'})$
- **negative** (goal opposite): $(\text{Bel } x\ p^{t'})^{t<t'}$ & $(\text{Goal } x\ \neg p^{t'})$
- **neutral**: $(\text{Bel } x\ p^{t'})^{t<t'}$ & $\neg(\text{Goal } x\ p^{t'})$ & $\neg(\text{Goal } x\ \neg p^{t'})$
- **ambivalent**: $(\text{Bel } x\ p^{t'})^{t<t'}$ & $(\text{Goal } x\ p^{t'})$ & $(\text{Goal } x\ \neg p^{t'})$

To be happy or to be a good predictor?

To be more subtle, given the Epistemic Goal that we have postulated in any true Expectation, one might say that in negative expectations always there is a minor conflict, since X on the one side desires, wishes that p [G1: (Goal x p)], but since she is induced (by some evidence or experience) to forecast that Not p, she also has the opposite goal [G2: (Goal x ¬p)]. However, this goal usually is not so relevant as the first objective, since it is just in order to confirm X to be a good predictor or that the world is predictable enough; it is just a by-product of control mechanisms and meta-goals. If the negative expectations result to be wrong, X is happy as for G1, but G2 is frustrated. Vice versa, if the negative expectation has been right, X is unhappy as for G1, but can have some 'comfort' because at least she is a good predictor, expert of the world. In positive expectations, since the G1 and G2 converge (that is X has the Goal that p both for intrinsic reasons, and for confirming her prediction and competence), when the prediction is wrong the frustration is appraised without compensation.

Expectations and Intentional (Goal-driven) Behavior

Intentional and in general goal-driven action requires and implies Expectations in strict sense, but not the other way around. Expectations are broader that intentional (or goal-directed) actions, they are not necessarily related to action; since even goals are not necessarily related to action.[8] First of all, there are Expectations also for goals we are not actively pursuing. Second, not all goals imply expectations. Inactive goals, or already realized goals, or discarded goals do not bring with them any expectation.

Expectation without Intention and Pragmatic Action

Only active and non-realized goals build Expectations. This covers two kinds of goals:

A) Active achievement goals [9]: goals to be achieved by the subject's action; to be brought about; it is not simply a matter of waiting for them.

B) Self-realizing achievement goals; the agent has nothing to do for achieving them (X has just to wait) since they are realized by other agents and she can just delegate (Castelfranchi 1997) this realization to them. The delegated 'agent' can either be "nature" and some natural process, and usually X can do nothing at all because the desired state only depends on the world ("tomorrow be a sunny day"; "to grow and become a woman"); or can be a social agent Y like X, acting in a common world. For example, Y stops the bus as desired by X, and X relies on this.

Having such a goal may perfectly produce an Expectation (positive or negative) when there also is a prediction about the desired event. X is just expecting, while doing nothing for realizing the Goal, but doing something for monitoring the world. If I wish that tomorrow will be sunny (since I

[7] We mention this because it is the case of proto-expectations or expectations in 'Anticipatory-Classifiers' based behaviors, strictly conceived as reactive (not really goal-driven) behaviors, but based on anticipatory representation of the outcomes.

[8] Although we are pushed – especially in English – to conceive 'goals' as 'objectives', 'targets' of some *action*.

[9] For a complete analysis we should also take into account the distinction between *achievement* and *maintenance* goals (see Castelfranchi 1997).

plan for a trip in the country) and I believe it (positive expectation: hope), I can do nothing for it being sunny, but when I wake up in the morning I check whether it is sunny or not. Let's call these 'passive expectation' while calling 'active expectations' those related to intentional pragmatic actions and active pursuit of the Goal. Obviously a passive expectation can become an active one during the evolution of the events.

Expectations in Intentions

As we said, no Intention is possible without Expectation, but this is not a new irreducible primitive, to be added for example in the BDI (Beliefs, Desires, Intentions) framework (Cohen & levesque 1990; Rao & Georgeff 1992). It can and must be recollected to beliefs and goals. And it is a molecule, not a set of atoms; a *mixed* attitude: in part epistemic, in part motivational.[10] In fact in order to deliberate to act and to commit to a given course of action (Bratman 1987) one should believe a lot of things (that it is to be preferred, that is not self-realizing or already realized, to have a plan, to be able and in condition for executing the actions, etc.). Among those beliefs supporting intentions (Castelfranchi 1996) some crucial ones are the beliefs about the expected effects of the actions (that motivated its choice) and the expected achievement. One cannot intend to do action μ in order to achieve p if she does not believe that after action μ is executed p will be true. Thus any Intention presupposes and entails a 'positive' Expectation.

More precisely, also a *weak* positive expectation is compatible with intentional behavior. At least one has not to believe that $\neg p$; otherwise her act would be completely irrational (subjectively useless). Thus there is a Weak Expectation, when X has the Goal (and in this case the Intention) that p and does not believes that not p in the future: $\neg\,(\text{Bel } x \neg (p^{t'}))^{t<t'}$ & $(\text{Goal } x\ p^{t'})$;

X is 'attempting', intentionally trying to realize p.

In any case in intentional action it is excluded a negative certain expectation

$(\text{Bel } x \neg (p^{t'}))^{t<t'}$ & $(\text{Goal } x\ p^{t'})$

We mean: acting with the certainty to fail. It would be fully irrational.

The quantitative aspects of mental attitudes and of their emergent configurations

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

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

P1: Beliefs & Goals have specific quantitative dimensions; that 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 (Galliers 1991).

Goals have a value, a subjective importance for the agent.

This gives us four extreme conditions (but in fact those variations are continuous and one should model precisely this continuity):

		Belief	
		High credibility (pretty sure)	**Low credibility** (perhaps)
Goal	**High value** (very important)	1	2
	Low value (marginal)	3	4

To simplify, we may have very important goals combined with uncertain predictions; pretty sure forecasts for not very relevant objectives; etc.

Thus, we should explicitly represent these dimensions of Goals and Beliefs:

$\text{Bel}^{\%}\ x\ p^{t}$; $\text{Goal}^{\%}\ x\ p^{t}$

Where % in Goals represents their subjective importance or value; while in Beliefs, % represents their subjective credibility, their certainty.

An Expectation (putting aside the Epistemic Goal) will be like this:

$$\boxed{\text{Bel}^{\%}\ x\ p^{t}\ \&\ \text{Goal}^{\%}\ x\ [\neg]\ p^{t}}$$

The subjective *quality* of those "configurations" or macro-attitudes will be very different precisely depending on

[10] In AI there have been other attempt to insert Expectations among the necessary mental ingredients of a BDI like agent (Corrêa & Coelho 1998). The difference is not only that we derive several "psychological" assumptions and consequences from our model, but also that we do not introduce Expectations as an additional primitive. We prefer to build these mental states on former ingredients (beliefs and goals/intentions) in order to have mental states that preserve both properties, epistemic and conative. Expectations have a specific functional role in practical reasoning that is better understood when those mental states are defined in a compositional fashion.

those parameters. Also the effects of the invalidation of an expectation are very different depending on:
a) the positive or negative character of the expectation;
b) the strengths of the components.

We also postulate that:

P2: The dynamics and the degree of the emergent configuration, of the Macro-attitude are strictly function of the dynamics and strength of its micro-components.

For example anxiety will probably be greater in box 2 than in 1, inferior in 4, nothing in 3. Box 2 (when the expectation is 'positive') produces an intense hope; and so on. Let us characterize a bit some of these emergent macro-attitudes.

Hope and fear

'Hope' is in our account (Miceli & Castelfranchi 2002; Castelfranchi & Lorini 2003) 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.*

$$\boxed{\text{Bel}^{\textbf{low}}\ x\ p^t\ \ \&\ \ \text{Goal}^{\textbf{high}}\ x\ p^t}$$

We may also have – it is true - '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.
Correspondingly one might characterize being afraid, 'fear', as an expectation of something bad, i.e. against our wishes:

$$\boxed{\text{Bel}^{\%}\ x\ p^t\ \ \&\ \ \text{Goal}^{\%}\ x\ \neg p^t}$$

but it seems that there can be 'fear' at any degree of certainty and of importance.[11]
Of course, these representations are seriously incomplete. We are ignoring their 'affective' and 'felt' component, which is definitely crucial. We are just providing their cognitive skeleton (Castelfranchi 2005).

Expecting Artificial-Agents

One reason for such a quite abstract, essential (and also incomplete) analysis is that this can be formalized and implemented for artificial creatures. Computers and robots can have different kinds of Expectations: low level perceptual expectations for monitoring the world; proto-intentions for monitoring the action and reinforcing it by learning; and high level explicit expectations. They are in fact able of making predictions on the physical world and on the other (also human) agents. They can do this on various bases (from inference and analogy to statistical learning, from laws and norms to mind reading and plan recognition) as we do; and they can have true 'purposive' behavior, intentional actions guided by pre-represented goals. Thus, they can entertain true Expectations. It would be necessary to also represent and use the strength and credibility of Beliefs (based on sources and evidences) (Castelfranchi 1996) and the value of the Goals (on which preferences and choices should be based). Given this and various kinds of Epistemic actions, one might model surprise, disappointment, relief, hope, fear, etc. in robots and software agents.
Which should be the advantage of having machines anxious like us?
Seriously speaking, we believe that these reactions (although unfelt and incomplete) would be very adaptive and useful for learning, for reacting, for interacting with the user and with other agents.

Analytical Disentanglement and the Gestalt character of mental attitudes

Moreover, a hard problem for symbolic (and analytic) cognitive science deserves to be underlined: *the mental Gestalt problem*. Disappointment, expectation, relief, etc. seem to be unitary subjective experiences, typical and recognizable "mental states"; they have a global character; although made up of (more) atomic components they form a *gestalt*. To use again the metaphor of molecules vs. atoms, the molecule (like 'water') has emergent and specific properties that its atoms (H & O) do not have. How can we account for this gestalt property in our analytic, symbolic, disentaglement framework? We have implicitly pointed out some possible solution to this problem. For example:
- A higher-level predicate exists (like 'EXPECT') and one can assume that although decomposable in and implying specific beliefs and goals, this molecular predicate is used by mental operations and rules.
- Or one might assume that the left part of a given rule for the activation of a specific goal is just the combined pattern: belief + goal; for example, an avoidance goal and behavior would be elicited by a serious negative *expectation* (and the associated 'fear'), not by the simple prediction of an event.

[11] 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; fear 'conative' and 'impulsive' aspect. 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?!") (Miceli & Castelfranchi in press). The more intense the felt fear, the more important the activate goal of avoidance (Castelfranchi 2005).

- One might assume that we "recognize" - or better "individuate" (and "construct")- our own mental state (thanks to this complex predicate or some complex rule) and that this "awareness" is part of the mental state: since we have a complex category or pattern of "expectation" or of "disappointment" we recognize and *have* (and feel) this complex mental state.

This would create some sort of "molecular" causal level. However, this might seem not enough in order to account for the gestaltic subjective experience, and reasonably something additional should be found in the direction of some typical "feeling" related to those cognitive configurations. Here we deal with the limits of any disembodied mind (and model).

The dynamic consequences of Expectations

As we said, also the effects of the *invalidation* of an expectation are very different 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 this changes her evaluation of and reaction to a given event?

Invalidated Expectations

We call invalidated expectation, an expectation that results to be wrong: i.e. while expecting that p at time t', X now beliefs that NOT p at time t'.

$(\mathrm{Bel}\ x\ p^{t'})^{t<t'} \leftarrow\rightarrow \mathbf{(Bel\ x\ \neg p^{t'})^{t''>t}}$

This crucial belief is *the 'invalidating'* belief.

- Relative to the goal component it represents "frustration", "goal-failure" (is the *frustrating* belief): I desire, wish, want that p but I know that not p.

FRUSTRATION: $(\mathrm{Goal}\ x\ p^{t'})$ & $(\mathrm{Bel}\ x\ \neg p^{t'})$

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

INVALIDATION: $(\mathrm{Bel}\ x\ p^{t'})^{t<t'}$ & $(\mathrm{Bel}\ x\ \neg p^{t'})^{t''>t}$

$(\mathrm{Bel}\ x\ p^{t'})^{t<t'}$ 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.* Given positive and negative Expectations and the answer of the world, that is the *frustrating* or *gratifying* belief, we have:

	P	¬P
Bel x p & Goal x p	no surprise + achievement	*surprise + frustration* ***disappointment***
Bel x ¬p & Goal x p	*surprise + non-frustration* ***relief***	no surprise + frustration

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:
$(\mathrm{Goal}^{\%}\ x\ p^{\ t'})^{t\ \&t'}$ & $(\mathrm{Bel}^{\%}\ x\ p^{\ t'})^{t}$ & $(\mathrm{Bel}^{\%}\ x\ \neg p^{\ t'})^{t'}$

At t X believes that at t' (later) 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 *sufferance* than simple frustration (Miceli & 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 (Bandura's "predictability" and related "controllability") and is disorienting; (iii) for the fact that losses of a pre-existing fortune are worst than missed gains (see below), and long expected and surely expected desired situation are so familiar and "sure" that we feel a sense of loss.

The stronger and well grounded the belief the more disorienting and restructuring is the *surprise* (and the stronger the consequences on our sense of predictability). The more important the goal the more *frustrated* the subject.

In Disappointment these effects are combined: *the more sure 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 function of both dimensions and components [12]. It seems to be felt as a unitary effect.

"How much are you disappointed?" "I'm very disappointed: I was sure to succeed"

"How much are you disappointed?" "I'm very disappointed: it was very important for me"

"How much are you disappointed?" "Not at all: it was not important for me"

"How much are you disappointed?" "Not at all: I have just tried; I was expecting 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 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

[12] 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.

"down" while expecting something bad, and now feel much better because this expectation was wrong.

RELIEF:
$(\text{Goal } x \neg p^{t'})$ & $(\text{Bel } x\ p^{t'})$ & $(\text{Bel } x \neg p^{t'})$[13]

- *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 treat, and the stronger the relief. The worry is already 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.

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. In fact expecting that a good state will continue is different from expecting that a good state (that currently is not real) becomes true; and it is different worrying about the cessation of a good state vs. worrying about the instauration of a bad event. Consequently, the Relief for the cessation of a painful state that X expected to continue is different from the Relief for the non-instauration of an expected bad situation. Analogously: the Disappointment for the unexpected non-prosecution of a welfare state (*loss*) is psychologically rather different from the non-achievement of an expected goal.

		FORECAST that P	
		Currently P (*expected continuation*)	**Currently Not P (*expected instauration*)**
ACTUALLY Not P	**GOAL P**	**1 Disappointment *Loss***	**2 Disappointment *Missed Gain***
	GOAL Not P	**3 Relief *Cessation, Alleviation***	**4 Relief *Escaped Danger***

More precisely (making constant the value of the Goal) the case of loss (1) is usually worst than (2), while (3) is better than (4). This is coherent with the theory of psychic suffering (Miceli & 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' (1), not just 'missed gains' (2), that is when the frustrated goal is a maintenance goal not an achievement goal.

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 Expectations.

For "implicit" beliefs we mean here a belief that is not 'written', contained in any 'data base' (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, while my knowledge that Buenos Aires is the capital city of Argentina is an explicit belief that I have in some memory and I have just to retrieve it, on the contrary my knowledge that Buenos Aires is not the capital city of Greece (or of Italy, or of India, or of …) is not in any memory, but can just be derived (when needed) from what I explicitly know. Until it remains implicit, merely potential, until is not derived, it has *no effect* in my mind; for example, 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"[14]. 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:

$$\boxed{\begin{array}{l}\text{Bel}^{\%}\ x\ p^{t}\\ \&\\ \text{Goal}^{\%}\ x\ p^{t}\end{array}} \quad \Rightarrow \quad \boxed{\begin{array}{l}\text{Bel}^{\%}\ x \neg p^{t}\\ \&\\ \text{Goal}^{\%}\ x\ p^{t}\end{array}}$$

This means that any hope implicitly contains some fear, and that any worry implicitly preserves some hope. But also means that when one get a 'relief' because a serious threat strongly expected is not arrived and the world is conforming to her desires, she also get (or can get) some exultance. It depends of her focus of attention and framing: is she focused on her worry and evanished treat, or on the unexpected achievement? Vice versa 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 implicit previous worry.

Not necessarily at the very moment that one feels a given emotion (for example fear) she also feels the complementary emotion (hope) in a sort of oscillation or ambivalence and affective mixture. Only when the belief is

[13] Or – obviously - (Goal x pt') & (Bel x ¬pt') & (Bel x pt').

[14] We are simplifying the argument. In fact it is possible that there is an interval of ignorance, some lack of evidences; that is that I 45% evaluate that p and 30% that Not p, having a gap of 25% neither in favor of p nor of Not p (Shafer 1976; Pezzulo *et al.* 2004).

explicitly represented and one can focus – at least for a moment – her attention on it, it can generate the corresponding emotion.

Concluding remarks

This analysis obviously is very simplistic, and reductionist. It misses a lot of important psychological aspects. As we mentioned, an important missed point is the fact that those mental states (especially when 'affective') are usually joined with bodily activation and feeling components, and these components –with their *intensity*- shape the whole subjective state and determine the nature of future reactions. Moreover, other cognitive aspects are elicited by and combined with those configurations. For example, in worrying the activity of monitoring, waiting, be more or less anxious. Now the degree of relief also depends on the presence and intensity of those somatic components and of those activities (Was the subject very stressed, feeling her stomach contracted? … Was she continuously checking and checking?) .

We also did not consider the important interaction between the two basic components and their strength. For example, there might be an influence of the goal on the belief. In 'motivated reasoning' (Kunda 1990), in wishful thinking we tend to believe more agreeable (goal conformable) beliefs and we defend ourselves from bad (goal opposite) beliefs. In Expectations we precisely have goal-related beliefs, thus – with an important value of the goal – we might be prone to go against the independent sources and evidences of our beliefs and change their credibility in conformity with their desirability. In other words, our predictions might be influenced by the value of the expected outcome. Vice versa, in some psychological attitude or personality one might reduce the concern, the value of the goal just in order to not feel so bad in case of failure, since she mainly focuses such an eventuality.

However, this simplification is just a necessary, preliminary step: nothing prevents AI and ALife from enriching this skeleton with more mussels and blood. This anatomy is necessary for identifying basic structural relationships between mental states, and – in this case- the crucial (sometimes hidden) role of expectations in mind.

Notice that –even with such a simplification - several nice predictions follow from this cognitive anatomy. For example, we predict that Disappointment implies Surprise, but not the other way around; or that Hope implies a Prediction, but not vice versa. We can predict that there is a contradiction between 'to be frightened of' something and be disappointed if it does not happen; or between forecasting that p and be surprised when it actually happens; or between 'hoping' that p and feeling down if it happens. We predict that a strong hope, when the prediction is realized, entails satisfaction, realization; while in the opposite case entails frustration, disappointment, and pain.

Will we have the satisfaction of surprising our artificial Agent, our computer or our domestic robot? And possibly even of disappointing them (as they frequently disappoint us)? We think so, and – as we said – this objective has been an additional reason for being schematic. Computers and robot can have Expectations and one might model robotic surprise, disappointment, relief, hope, fear, etc.

Of course, to really having artificial fear or hope one should reproduce or simulate also the 'affective' component, that is the 'feeling', by providing to computers, artificial agents, and robots a 'body' not simply a hardware. This means introducing some form of proprioception and enteroception, pain and pleasure, feeling what happens to the body and its internal states and events, its automatic reactions to the world; and modeling the impact of these signals (*motions*) on the 'mental' representations and activity (Castelfranchi 2005). This is still quite far to be achieved. This is why we can have for the moment only the 'cold' counterpart of those affective states, just reduced to the mental representations on which they are based.

However, the objective remains that of building some (useless?) anxious machine.

Acknowledgement

I would like to thank Maria Miceli and Emiliano Lorini that have been working with me on these topics for several years, and all the *MindRaces* ISTC group for precious discussions. This research has been funded by the European Project *MindRaces* N°511931- EC's 6th Framework Programme, IST priority- Unit: Cognitive Systems (http://www.mindraces.org/) and the European HUMAINE Network (http://emotion-research.net/).

References

Bandura A., (1990) Self-efficacy mechanism in human agency. *American Psychologist*, 37, 122-147.

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

Bickhard, M. H. (2005) Anticipation and Representation. In Proceedings of *From Reactive to Anticipatory Cognitive Embodied Systems*, AAAI Fall Symposium Series Technical Report.

Butz, M.V. (2002) *Anticipatory learning classifier system* Boston, MA: Kluwer Academic Publisher.

Butz, M.V. and Hoffman, J. (2002). Anticipations control behavior: Animal behavior in an anticipatory learning classifier system. *Adaptive Behavior*, 10, 75-96.

Castelfranchi, C. (2005) *Ri-emboding 'hope' and 'fear'* T.R. European Project HUMAINE, (2005).

Castelfranchi, C., (1997) Individual social action. In G. Holmstrom-Hintikka and R. Tuomela (eds.), *Contemporary theory of action*, vol.II, Kluwer, Dordrecht, pp. 163-192.

Castelfranchi, C., (1996) Reasons: Belief Support and Goal Dynamics. *Mathware & Soft Computing, 3*, 233-47.

Castelfranchi, C. and Lorini, E. (2003) Cognitive Anatomy and Functions of Expectations. In *Proceedings of IJCAI'03 Workshop on Cognitive Modeling of Agents and Multi-Agent Interactions*, Acapulco, Mexico.

Cohen, P. R. and Levesque, H. J., (1990) Intention is choice with commitment. *Artificial Intelligence, 42*, 213-261.

Cooper, J. and Fazio, R. H., (1984) A new look at dissonance theory. In L. Berkovitz (Ed.), *Advances in experimental social psychology, Vol. 17*, pp. 229-266, San Diego, CA: Academic Press.

Corrêa, M. and Coelho, E. (1998) Agent's programming from a mental states framework. In Proceedings of the *14th Brazilian Symposium on Artificial Intelligence (SBIA98),* Lecture Notes in AI 1515, pp. 31-39, Springer-Verlag.

Drescher, G. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. Cambridge: MIT Press.

Galliers, J.R. (1991) Modelling Autonomous Belief Revision in Dialogue, In *Decentralized AI-2*, Y. Demazeau, J.P. Mueller (eds), pp. 231-43. Armsterdam: Elsevier.

Jordan, M. I. and Wolpert, D. M. (1999) Computational Motor Control. In M. Gazzaniga (Ed.), *The Cognitive Neuroscience,* Cambridge, MA: MIT Press.

Kahneman, D. and Miller, D. T., (1986). Norm Theory: Comparing reality to its alternatives. *Psychological Review*, 93,136-153.

Kirsh, D. and Maglio. P. (1994) On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*, 513-549.

Kunda, Z., (1990). The case of motivated reasoning. *Psychological Bullettin, 108*, 480-498.

Lorini, E. and Castelfranchi, C.(in preparation). *Towards a cognitive model of Surprise*.

Lorini, E. and Castelfranchi, C., (2004) The role of epistemic actions in expectations. In *Proceedings of Second Workshop of Anticipatory Behavior in Adaptive Learning Systems (ABIALS 2004)*.

Miceli, M. and Castelfranchi,, C. (in press) For a Cognitive Theory of Anxiety. *British Medical Journal: Anxiety disorder*.

Miceli, M. and Castelfranchi, C. (2002). The Mind and the Future. The (Negative) Power of Expectations. *Theory & Psychology*, 12(3), 335-366.

Miceli, M. and Castelfranchi, C. (2000) The role of evaluation in cognition and social interaction. In K. Dautenhahn (Ed.), *Human cognition and agent technology*. Amsterdam: Benjamins, pp. 225-61.

Miceli, M. and Castelfranchi, C. (1997). Basic principles of psychic suffering: A preliminary account. *Theory & Psychology*, *7*, 769-798.

Miller, G., Galanter, E., Pribram, K. H. (1960) *Plans and the structure of the behavior*. Rinehart & Winston, New York.

Ortony, A. and Partridge, O. (1987) Surprisingness and expectation failure: What's the difference? In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 106-108, Los Altos, CA: Morgan Kaufmann.

Pezzulo, G., Lorini, E., Calvi G., (2004). How do I know how much I don't know? A cognitive approach about Uncertainty and Ignorance. In *Proceedings of 26th Annual Meeting of the Cognitive Science Society (CogSci 2004),* Chicago, USA.

Rao, A. S. and Georgeff, M. P., (1992) An abstract architecture for rational agents. In Proceedings of the *Third International Conference on Principles of Knowledge Representation and Reasoning*, C. Rich, W. Swartout, and B. Nebel (Eds.), pp. 439-449, Morgan Kaufmann Publishers, San Mateo, CA.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press, Cambridge.