The cognitive structure of Surprise: looking for basic principles

Emiliano Lorini, Castelfranchi Cristiano

The cognitive structure of surprise: looking for basic principles

Emiliano Lorini and **Cristiano Castelfranchi**

12th September 2006

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

Abstract

We develop a conceptual and formal clarification of the notion of *surprise* as a belief-based phenomenon by exploring a rich typology. Each kind of surprise is associated with a particular phase of the cognitive processing and involves particular kinds of epistemic representations (representations and expectations under scrutiny, implicit beliefs, presuppositions). We define two main kinds of surprise: *mismatch-based surprise* and *astonishment*. In the central part of the paper we suggest how a formal model of surprise can be integrated with a formal model of *belief change*. We investigate the role of surprise in triggering the process of belief reconsideration. There are a number of models of surprise developed in psychology of emotion. We provide several comparisons of our approach with those models.

1 Introduction

Surprise is the automatic reaction to a mismatch. It is a (felt) reaction/response of alert and arousal due to an inconsistency (discrepancy, mismatch, non-assimilation, lack of integration) between an incoming input and our previous knowledge, in particular an actual prediction or a potential prediction. It invokes and mobilizes resources at disposal of an activity for a better epistemic processing of this strange information (attention, search, belief revision, etc.), but also for coping with the potential threat. Surprise is 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.
Moreover, there are different kinds and levels of Surprise. There is a first-hand surprise - the most peripheral one, just due to the perceptual mismatch between what the agent *sees* and its sensory-motor expectations; while 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 and explanation of the perceived data (Meyer et al., 1997). This is surprise due to the implausibility of the new information. 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 and Miller, 1986) or on mental simulation. On the other hand 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 likes Mary, so he will invite her for a dinner; He decided to go in vacation, so he will not be here on Monday").

In this work we are mainly interested in the analysis of those forms of Surprise which involve symbolic high-level representations of expected events or objects and *a recognized event or object* in the external world. We are not going to analyze those forms of surprise due to the mismatch between low-level sensory expectations and a raw perceptual input (raw sensor datum). We restrict our analysis to those forms of cognitive surprise involving an already *perceived and recognized* object or event.[1] In order to account for the process of *cognitive recognition* we have developed in our complementary work (Lorini and Castelfranchi, 2006b) an *abduction-based procedure of explanation assessment and selection*. This procedure has the function of returning the *best explanation* of the data obtained by the sensors. We have shown in Lorini and Castelfranchi (2006b) that this selected explanation can mismatch with some pre-existent cognitive representation and therefore be responsible for the generation of surprise. In this work we do not introduce any *abduction-based procedure of explanation selection* and we simply assume that an agent can directly perceive and recognize an object or event without interpreting the perceptual raw sensor data by means of some abductive procedure.

In section 2 a formal logic of beliefs and probabilities is introduced. This simple logic is developed in order to provide formal representations of several kinds of mental attitudes. We provide definitions for beliefs and expectations of an agent. Moreover we characterize the notion of *scrutinized expectation*, i.e. the expectation on which the agent is focusing its attention and that the agent tries to match with the perceptual data. We introduce: 1) the special kind of mental operation *retrieve* with the function of introducing a new expectation into the mental *Test* (scrutiny) *space* of the agent; 2) the special action of *perceiving* some fact with the function of modifying the agent's *perceptual data*.

In section 3 we analyze two different kinds of surprise which involve all informational mental states characterized in section 2. We argue that these forms of surprise are the basic forms of surprise in cognitive systems involving symbolic high-level representations of expected events.[2]

[1] The necessity for a distinction between a mere activity of *seeing*, *hearing*, *smelling* something and a *complex cognitive* activity of *perceptual recognition* of an object or event has also been stressed by Dretske (1981).

[2] In the extended version of this paper (Lorini and Castelfranchi, 2006a) we investigate also those forms of surprise due to the invalidation of the agent's presupposed frame. We provide a general definition of frame (or script) as agglomerate of conditional beliefs and argue that a special kind of surprise (called *disorientation*) arises from the invalidation and revision of the conditional beliefs which are part of a given presupposed frame of the agent.

1. *Mismatch-based surprise (given the conflict between a perceived fact and a scrutinized representation*). I'm actively checking whether a certain event is happening, that is I have an endogenous anticipatory explicit representation of the next input and I attempt to match the incoming data against it. If there is a mismatch (conflict) between the two representations there is surprise. The intensity of this form of surprise is a function of the probability assigned to the expectation conflicting with the perceived fact.

2. *Astonishment* or *surprise in recognition*. I perceive a certain fact and recognize the implausibility of this. The recognition of implausibility of the perceived fact can based on two different kinds of mental processes.

   (a) On one side, after perceiving a certain fact φ that I was not actively expecting, I can retrieve from my background knowledge the probability of the event and conclude that "I would not have expected that event". The intensity of astonishment is a function of the probability assigned to the negation of the perceived fact ($\neg\varphi$).

   (b) On the other side after perceiving a certain fact φ I infer from my explicit beliefs the negation of the perceived fact.

We will argue that the previous typology of Surprise is based on the characterization of different kinds of informational mental states.[3] The following figure 1 summarizes them.
According to our view an agent has a *representation under scrutiny* (a focused expectation) and this must be distinguished from all those accessible *representations and expectations in background* (at an unconscious and automatic level). This distinction between *expectations and beliefs under scrutiny* and *background expectations and beliefs* looks similar to Kahneman & Tversky's distinction (Kahneman and Tversky, 1982) between *active expectations* and *passive expectations*. According to Kahneman & Tversky the former occupy consciousness and draws on the limited capacity of attention; the latter kind are available at a mere automatic and effortless level. Passive expectations could be the product of *priming*. Moreover, an agent looks at the world and acts in the world within the assumption of a presupposed complex mental framework, of a given *presupposed frame* (or script) which represents its unproblematic interpretation of the context of the situation where its action and perception are situated and which supports all (focused and background) expectations. Thus when presupposing to be entered in a restaurant the agent can reasonably expect to perceive a waiter, some tables and so on...
Expectations and beliefs under scrutiny, *background expectations and beliefs* and *presupposed frame* are members of the general set of *explicit informational mental states*. The last category of informational mental states is the category of *implicit expectations and beliefs*, that is all those (potential) beliefs and expectations that can be inferred from explicit beliefs and expectations (Ortony and Partridge, 1987; Levesque, 1984).

[3] We use the term "informational mental state" in order to distinguish it from a "motivational mental state" (a desire, intention, wish, goal and so on).

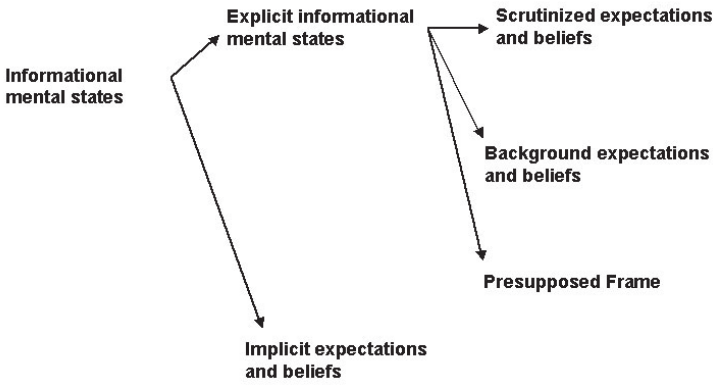

Figure 1: Ontology of informational mental states

In section 4 we try to build a bridge between our model of surprise and the theory of belief change. We start by extending the formal language and introducing *update processes* whose function is the modification of beliefs and expectations of the agent after the perception of a new fact. We extensively investigate the role of surprise in triggering belief change and we provide cognitive principles governing this kind of cognitive phenomenon.

Let us sum up here the major claims of our work.
Surprise is a very relevant belief-based phenomenon, with mental and behavioral aspects. In order to understand it, it is necessary to model the relationships between basic properties of beliefs (like their strength, their being explicit or implicit, their being passively assumed or actively tested) and surprise kinds and dimensions. It is a belief-based phenomenon because it is based on an actual or potential prediction formulated on the basis of the other beliefs, and because one of its main effect is the revision of our assumptions: the new data must be assimilated in our knowledge base, and our beliefs (base of our wrong prediction) must be revised.
We claim that in the literature there are very good studies and claims on surprise: like the idea that it depends on expectations, or the claim that its intensity depends on the "unexpectedness" of the stimulus (Ortony and Partridge, 1987; Meyer et al., 1997), the claim that one can deal with this in terms of information or probability, or the claim that there are different kinds of expectations - active versus passive, explicit versus implicit - which produce different kinds of surprise. However, we claim that more careful distinctions and clear characterizations of surprise are needed. We present here a 'theory driven' account of surprise, an analytical cognitive model which allows us to predict

and distinguish different levels and kinds of surprise, not necessarily already discriminated in the empirical researches. Sometimes even common sense concepts look much richer: for example, the concept of "astonished" is not identical to the concept of "surprised", or to the concept of "being disoriented". We try to provide some principled and precise distinctions of these different levels and kinds of surprise and to formalize some relevant properties of them. It is not for example the same kind of surprise when we immediately recover from it, while saying "What a stupid I am! It is obvious! I should have expected this", and when we remain in a strong and long suspended situation, unable to realize/accept and understand what has happened. Of course, we take into account some important psychological models (Ortony and Partridge, 1987; Meyer et al., 1997; Reisenzein et al., 1996), which are very relevant and interesting, and also currently implemented models in AI (Macedo and Cardoso, 2001). But they are still quite poor and simplified. They for example do not clearly distinguish between the surprise relative to the invalidation of a strong anticipated expectation, and the surprise relative to the degree of "unexpectedness" of the new incoming data. The two processes are - in our model - related and partially complementary, but not identical.
A complex view of surprise and of its nature and functions is necessary to understand the phenomenon. We do not model all its aspects. We do not investigate the experiential, phenomenal character of surprise (Reisenzein, 2000): surprise as a felt signal.[4] Moreover we do not model functional aspects of surprise (alert, learning, etc.) except those related to belief reconsideration (Section 4). As some psychologists have stressed (Meyer et al., 1991, 1997) surprise can culminate in a process of belief change. In this work we want to try to suggest some interesting ways a formal and computational model of *surprise* viewed as a belief-based phenomenon can be integrated with a formal model of *belief change*. Indeed belief revision theory has been mostly focused on the problem of finding general principles characterizing the process of belief change, but has completely neglected to account for the causal precursors of this kind of process.

2 Formal bases

We define in this section the formal language with the related syntax and semantic. We use a logic of probabilistic quantified beliefs with a semantics similar to the semantics given in Fagin and Halpern (1994); Halpern (2003). We add to the basic formal language the standard dynamic operator for talking about *actions*. Moreover, we use special formal constructs to denote the *representation under scrutiny (or under test)* of a given agent and the agent's *perceptual data* collected by its sensors.
We characterize two special kinds of actions: the mental operation *retrieve* which has the function of moving new information into the *scrutiny (test) space*; - the action *perceive* which has the function of modifying the agent's perceptual data. The main function of the formalism is to disambiguate the relevant concepts and notions of our model of surprise.

[4]Notice that "to feel surprised" should not be confused with "having awareness of our own surprise" (for a distinction between *phenomenal consciousness* and *access consciousness* see Bloch, 1995; Chalmers, 1995).

2.1 Syntax

The primitives of the formal language are the following:

- A set of *atomic actions* $AT = \{a, b, ...\}$.
- A set of propositional variables $\Pi = \{p, q, ...\}$.

The set $PROP = \{\varphi, \psi, ...\}$ is the set of *propositional formulas* defined by the closure of Π under the Boolean operations $\wedge$ and $\neg$. On one hand OBS is the set of *perceptual actions* defined as the smallest set such that:

- if $\varphi \in PROP$ then $observe(\varphi) \in OBS$.

On the other hand $RETR$ the is the set of *retrieve mental operations* defined as the smallest set such that:

- if $\varphi \in PROP$ then $retrieve(\varphi) \in RETR$.

$ACT = \{\alpha, \beta, ...\}$ is the set of *actions* which is defined as the smallest set such that:

- $AT \subseteq ACT$;
- $OBS \subseteq ACT$;
- $RETR \subseteq ACT$;
- if α and $\beta \in ACT$ then $\alpha; \beta \in ACT$ (sequential composition).

Our language $\mathcal{L}_{\mathcal{SURP}}$ is given by the following rule in extended Backus-Naur Form:

$$\Phi ::= p|\neg\Phi|\Phi_1 \wedge \Phi_2|Bel\Phi|\,[\alpha]\,\Phi|d_1P(\Phi_1)+...+d_nP(\Phi_n) \geq c|Test(\varphi)|Datum(\varphi)$$

where $p \in \Pi$, $\varphi \in PROP$, $\alpha \in ACT$ and $d_1, ..., d_n, c$ are real numbers. A *primitive term* is a an expression of the form $P(\Phi)$. A *basic probability formula* is a statement of the form $P(\Phi) \geq c$. A *term* is an expression of the form $d_1P(\Phi_1) + ... + d_nP(\Phi_n)$ where $d_1, ..., d_n$ are real numbers and $n \geq 1$. Finally a *probability formula* is a statement of the form $t \geq c$ where *t* is a *term* and *c* is a *real number*. We call formulas of the form $Bel\Phi$ *belief formulas*, formulas of the form $Test(\varphi)$ *test formulas* and formulas of the form $Datum(\varphi)$ *perception formulas.* $Bel\Phi$ reads "the agent believes that Φ"; $P(\Phi) \geq c$ reads "the agent assigns to the fact φ at least probability *c*"; $[\alpha]\,\Phi$ reads "always if the agent performs action α then Φ holds after α's occurrence"; $Test(\varphi)$ reads "φ is the representation that the agent is scrutinizing"; $Datum(\varphi)$ reads "φ is a *datum* perceived by the agent".
Propositional formula φ such that $Test(\varphi)$ should be considered the content of the expectation that the agent is actually scrutinizing and comparing and matching with the incoming input data. On the other hand propositional formula φ such that $Datum(\varphi)$ should be considered a *datum* obtained by the agent's sensors. With *perceptual datum* we mean here something similar to the notion of *datum* given in Rescher (1976). A *perceptual datum* is in our vocabulary some piece information gathered by the agent's

sensors which is a candidate for becoming a belief of the agent.[5] It will be shown below that both *perceptual data* and *scrutinized representations* play a crucial role within the surprise processing.
Moreover we use the following abbreviations:

$$\langle\alpha\rangle\,\Phi =_{def} \neg\,[\alpha]\,\neg\Phi;$$

$$\textstyle\sum_{i=1}^{n} d_i P(\Phi_i) \geq c =_{def} d_1 P(\Phi_1) + ... + d_n P(\Phi_n) \geq c$$

$$d_1 P(\Phi_1) \geq d_2 P(\Phi_2) =_{def} d_1 P(\Phi_1) - d_2 P(\Phi_2) \geq 0$$

$$\textstyle\sum_{i=1}^{n} d_i P(\Phi_i) \leq c =_{def} \sum_{i=1}^{n} -d_i P(\Phi_i) \geq -c$$

$$\textstyle\sum_{i=1}^{n} d_i P(\Phi_i) < c =_{def} \neg(\sum_{i=1}^{n} d_i P(\Phi_i) \geq c)$$

$$\textstyle\sum_{i=1}^{n} d_i P(\Phi_i) > c =_{def} \neg(\sum_{i=1}^{n} d_i P(\Phi_i) \leq c)$$

$$\textstyle\sum_{i=1}^{n} d_i P(\Phi_i) = c =_{def} \sum_{i=1}^{n} d_i P(\Phi_i) \leq c \wedge \sum_{i=1}^{n} d_i P(\Phi_i) \geq c$$

2.2 Semantics

We define with **M** the class of models of the form $M =< W, B, R_0, R_1, R_2, P, TEST, DATA, \pi >$ where each element of the tuple is defined as follows.

- $W = \{w, w', w'', ...\}$ is a non-empty set of possible worlds.
- B is a mapping $B : W \longrightarrow 2^W$ associating a set of possible world $B(w)$ to each possible world *w.* The elements in $B(w)$ are the alternative (worlds) that the agent considers possible at world *w*.
- R_0, R_1, R_2 are mapping

  1. $R_0 : AT \longrightarrow (W \longrightarrow 2^W)$
  2. $R_1 : OBS \longrightarrow (W \longrightarrow 2^W)$
  3. $R_2 : RETR \longrightarrow (W \longrightarrow 2^W)$

  associating sets of possible worlds $R_0^a(w)$, $R_1^{observe(\varphi)}(w)$ and $R_2^{retrieve(\varphi)}(w)$ to each possible world *w*. Those worlds *w'* such that $w' \in R_0^a(w)$, $w' \in R_1^{observe(\varphi)}(w)$ and $w' \in R_2^{retrieve(\varphi)}(w)$ are respectively those worlds which are achievable from *w* by doing the atomic action *a*, achievable by doing the action of perceiving φ and achievable by doing the operation of retrieving the expectation that φ from the background level.

[5] The need for the distinction between *data* and *beliefs* has been addressed by several other authors (see Paglieri, 2004 on this). For instance Tamminga (2001) advocated the need for two levels of explanation in dealing with belief revision, namely *information* (*data*) and *beliefs*. This leads him to describe belief revision as a two steps process: first, information revision, managed by applying a paraconsistent monotonic logic of first-degree entailment; second, belief extraction, that takes care of assuring nonmonotonicity, consistency, and closure under logical consequence. In Tamminga's work, the main focus is placed on inconsistency at the level of information *(*data*)* vs. consistency at the level of beliefs.

- P is a function which associates with each world w in W a probability space $P(w) = (W_w, X_w)$ where:
  - $W_w \subseteq W$ is called *sample space*;
  - X_w is a *probability function* defined on W_w such that $X_w : W_w \longrightarrow [0, 1]$ and

$$\forall w \in W \sum_{w' \in W_w} X_w(w') = 1$$

- *TEST* is a (test) function $TEST : W \longrightarrow PROP$[6] which assigns a *propositional formula* to each possible world. This function returns the *representation* that the agent is scrutinizing at a certain world, i.e. the representation on which the agent is focusing its attention and that the agent matches with the perceptual data.
- *DATA* is a (perception) function $DATA : W \longrightarrow PROP$ which assigns a *propositional formula* to each possible world. The function returns the *datum* obtained by the agent 's sensors at world *w*.
- $\pi : \Pi \longrightarrow 2^W$ assigns a set of worlds to each propositional variable.

Here we suppose that *B, TEST, DATA,* every R_0^a, every $R_1^{observe(\varphi)}$ and every $R_2^{retrieve(\varphi)}$ are partial functions.
We use the following two notational abbreviations:

- (Domain): $||\Phi||^{W_w} = \{w' \in W_w | M, w' \vDash \varphi\}$;
- (Probability of a Domain): $X_w(||\Phi||^{W_w}) = \sum_{w' \in ||\Phi||^{W_w}} X_w(w')$.

Truth conditions

- $M, w \vDash p \Longleftrightarrow w \in \pi(p)$
- $M, w \vDash \neg\Phi \Longleftrightarrow$ not $M, w \vDash \Phi$
- $M, w \vDash \Phi_1 \wedge \Phi_2 \Longleftrightarrow M, w \vDash \Phi_1$ and $M, w \vDash \Phi_2$
- $M, w \vDash Bel\Phi \Longleftrightarrow \forall w'$ if $w' \in B(w)$ then $M, w' \vDash \Phi$
- $M, w \vDash d_1 P(\Phi_1) + ... + d_n P(\Phi_n) \geq c \Longleftrightarrow d_1 X_w(||\Phi_1||^{W_w}) + ... + d_n X_w(||\Phi_n||^{W_w}) \geq c$
- $M, w \vDash Test(\varphi) \Longleftrightarrow \varphi = TEST(w)$
- $M, w \vDash Datum(\varphi) \Longleftrightarrow \varphi = DATA(w)$
- $M, w \vDash [\alpha]\,\Phi \Longleftrightarrow \forall w'$ if $w' \in R^\alpha(w)$ then $M, w' \vDash \Phi$

where $R^\alpha(w)$ is defined according to the following (1), (2), (3) and (4).

[6] Our *test function* is comparable to the *awareness* function defined in Fagin and Halpern (1987).

1. $R^a(w) = R_0^a(w)$;
2. $R^{observe(\varphi)}(w) = R_1^{observe(\varphi)}(w)$;
3. $R^{retrieve(\varphi)}(w) = R_2^{retrieve(\varphi)}(w)$;
4. $R^{\alpha;\beta}(w) = (R^\beta \circ R^\alpha)(w)$.

2.3 Basic properties and definitions

For what concerns the probabilistic fragment of our logic we inherit all axioms and inference rules given in Halpern (2003); Fagin and Halpern (1994). Soundness and completeness of this deductive system for a logic of belief and probability has been proved. The axiom system is made of three different kinds of axioms and inference rules. Axioms and inference rules are given for: 1) propositional reasoning and *Bel* modal operator[7]; 2) for reasoning about probability[8]; 3) for reasoning about linear inequalities[9].
Moreover we suppose here that believing that Φ holds implies that the maximum value of probability is assigned to Φ. Formally:

$(Incl_{Bel/Prob})\ Bel\Phi \rightarrow (Prob(\Phi) = 1)$.

This axiom requires that the set of worlds which are considered possible by the agent in an arbitrary world *w* is a superset of the *sample space* with respect to the arbitrary world *w*:

- for every $w \in W\ W_w \subseteq B(w)$.

With respect to the dynamic component we use standard axioms from dynamic logic. We take the axioms and inference rules of the basic normal modal logic for the dynamic operator and the standard axiom for sequential composition:

$(K)\ [\alpha]\,(\Phi \rightarrow \Psi) \wedge [\alpha]\,\Phi \rightarrow [\alpha]\,\Psi$
$([\alpha] - Necessitation)$ From $\vdash \Phi$ infer $\vdash [\alpha]\,\Phi$
$(Composition)\ [\alpha]\,[\beta]\,\Phi \longleftrightarrow [\alpha;\beta]\,\Phi$.

Moreover we suppose the following axioms for *atomic actions* and *perceptual actions*:

$(Det_{At})\ \langle a\rangle\,\Phi \rightarrow [b]\,\Phi$
$(Perc_1)\ \varphi \longleftrightarrow \langle observe(\varphi)\rangle\,\top$
$(Perc_2)\ [observe(\varphi)]\,Datum(\varphi)$

[7] a) All instances of propositional tautologies; b) Modus ponens: from $\vdash \Phi$ and $\Phi \rightarrow \Psi$ infer Ψ; c) K-axiom for *Bel*: $Bel(\Phi \rightarrow \Psi) \wedge Bel\Phi \rightarrow Bel\Psi$; d) *Bel*-Necessitation: From $\vdash \Phi$ infer $\vdash Bel\Phi$.

[8] a) Nonnegativity: $P(\varphi) \geq 0$; b) Probability of Truth: $P(\top) = 1$; c) Additivity: $P(\Phi_1 \wedge \Phi_2) + P(\Phi_1 \wedge \neg\Phi_2) = P(\Phi_1)$; d) Equivalence: From $\vdash \Phi_1 \longleftrightarrow \Phi_2$ infer $\vdash P(\Phi_1) = P(\Phi_2)$.

[9] See Fagin and Halpern (1994).

$(Perc_3)$ $\langle observe(\varphi)\rangle\, \Phi \rightarrow [observe(\varphi)]\, \Phi$.

$(Perc_1)$ says that: 1) it is always possible for the agent to perceive φ if φ is true in the external world and 2) if it is possible for the agent to perceive φ then φ is true in the external world. $(Perc_2)$ says that after φ is perceived, φ becomes a *perceptual datum* that is the action of *perceiving* φ moves a new datum φ into the *data space* of the agent. $(Perc_3)$ guarantees that *perceptual actions* are deterministic. (Det_{At}) guarantees that *atomic actions* follow the same path.
We note that the previous axioms correspond to the following semantic constraints:

- for every $w \in W$ if $w' \in R_0^a(w)$ and $w'' \in R_0^b(w)$ then $w' = w''$;
- for every $w \in W$ $R_1^{observe(\varphi)}(w) \neq \emptyset$ if and only if $M, w \vDash \varphi$;
- for every $w \in W$ if $w' \in R_1^{observe(\varphi)}(w)$ then $\varphi = DATA(w')$;
- for every $w \in W$ if $w' \in R_1^{observe(\varphi)}(w)$ and $w'' \in R_1^{observe(\varphi)}(w)$ then $w' = w''$.

Finally we suppose that the following are valid properties of *retrieve mental operations*:

$(Retr_1)$ $\langle retrieve(\varphi)\rangle \top \rightarrow \neg Test(\varphi)$
$(Retr_2)$ $[retrieve(\varphi)]\, Test(\varphi)$
$(Retr_3)$ $\langle retrieve(\varphi)\rangle\, \Phi \rightarrow [retrieve(\varphi)]\, \Phi$.

$(Retr_1)$ says that if it is possible for the agent to retrieve φ then φ is a a representation which is not actually scrutinized. $(Retr_2)$ says that after φ gets retrieved, φ is scrutinized by the agent. Thus the mental operation of *retrieving* φ has the function of modifying the mental setting of the agent, by moving a new representation φ into the *test space* of the agent. Finally $(Retr_3)$ guarantees determinism for *retrieve mental operations*. We note that the previous axioms correspond to the following semantic constraints:

- for every $w \in W$ if $\varphi = TEST(w)$ then $R_2^{retrieve(\varphi)}(w) = \emptyset$;
- for every $w \in W$ $w' \in R_2^{observe(\varphi)}(w)$ then $\varphi = TEST(w')$;
- for every $w \in W$ if $w' \in R_2^{retrieve(\varphi)}(w)$ and $w'' \in R_2^{retrieve(\varphi)}(w)$ then $w' = w''$.

We call *SURPRISE* the logic axiomatized by the axioms and inference rules for probabilities and beliefs given in Halpern (2003); Fagin and Halpern (1994) and discussed above, the axiom $Incl_{Bel/Prob}$, the previous axioms and inference rules for actions in general and the special axioms for *atomic actions*, *perceptual actions* and *retrieve mental operations*. We call *SURPRISE models* the set of models $\mathbf{M}_{Surp} \subseteq \mathbf{M}$ satisfying all the semantic constraints imposed in this section and write $\models_{Surp} \varphi$ if φ is valid in all *SURPRISE models.* Moreover we write $\vdash_{Surp} \varphi$ if φ is a theorem of *SURPRISE*.

Having defined *retrieve mental operation* and formulated their properties we can characterize the notion of background expectation (or background belief). A background (or passive) expectation is in our vocabulary an expectation whose content is available and accessible by means of a *retrieve mental operation*, that is a background expectation is an expectation whose content can be mentally retrieved. Formally:

$Background(\varphi) =_{def} \langle retrieve(\varphi)\rangle \top.$

The present distinction between expectations and beliefs under scrutiny of the form $Test(\varphi)$ and background expectations and beliefs of the form $Background(\varphi)$ looks similar to the distinction given in psychology between *active expectations* and *passive expectations* (Tversky and Koehler, 1994; Kahneman and Tversky, 1982). According to Kahneman & Tversky the former "occupy consciousness and draws on the limited capacity of attention"; the latter kind are "automatic and effortless". Passive expectations can be either permanent, such as categories and assumptions about the external world, or temporary, such as the *priming effects* in psychological experiments.[10]

3 Kinds of Surprise

3.1 Mismatch-based surprise and astonishment

It is the objective of having an operational and cognitively plausible model of surprise which gives rise to the need to introduce and exploit the notion of *representation under scrutiny* (*representation to be tested*). Indeed we want to model realistic cognitive systems which process input data and which are focused on a small portion of their internal information state. The purpose of this section is to clarify the distinction between *mismatch-based surprise* (there is a recognized conflict between the agent's input data and the agent's representation under scrutiny) and *astonishment*. While the notion of *mismatch-based surprise* is an operational notion and is associated with a recognized logical conflict between the incoming information and a representation under scrutiny, *astonishment* is in our view the response to the recognized implausibility of the input data. When I am astonished about something, I cannot believe what I see and this presupposes that I'm trying to believe, I'm trying to find an explanation for what I see, but I'm suspended. *Astonishment* seems to be due to a difficulty, to a delay due to this process of integration, of accounting for, which in this case is not automatic and fast, not immediately successful. We cannot in fact believe something just putting it in our belief base; we must check about consistency (especially if there are reasons for suspecting some inconsistency). If the actual input generates an intense astonishment then it means that the input is unexpected and rather unpredictable from my actual beliefs. If I have to accept it, I have to adjust my beliefs in such a way that they can account for this unexpected event. Generally in order to cope with an intense astonishment, I need a deep and large revision of my well consolidated beliefs.

[10] Several empirical evidences exist showing that in being active and available at an automatic and effortless level background (passive) expectations can affect subject's performances and judgments and can conflict with conscious (scrutinized) expectations (on this see Matt et al., 1992; Sommer et al., 1998).

Example 1. Consider a person being in a terrible delay. He needs to take a train from Florence to Rome at 8:00 a.m. It is 7:56 a.m. and the guy is still running to reach the Florence station. Finally he arrives at the station at exactly 8:00 a.m. and checks whether *the train for Rome is standing in the station*. At the moment of the perceptual test the agent has the representation of the *train for Rome standing in the station* explicit in his mind and attributes a high probability to this fact. When the agent perceives *the train for Rome in not standing* the agent gets very surprised since the incoming representation (logically) conflicts with the explicit representation of *the train for Rome is standing in the station* and the probability assigned to the fact *the train for Rome is standing in the station* is very high. This kind of Surprise is what we call *mismatch-based surprise*.

Example 2. It is 5:50 p.m. and Bill is working in his office when Mary phones Bill and tells him: "I will come to your office at 6 p.m.! Wait for me there!". After Mary's call Bill decides to stop working and to rest until Mary will arrive. Bill expects with high probability that *Mary will knock on the door of the office at 6 p.m.* and focuses his attention on this. It is 5:53 p.m. and suddenly someone knocks on the door. Bill opens the door and sees that a policeman is standing in front of the door. There is not logical conflict between the scrutinized representation *Mary will knock on the door of the office at 6 p.m.* and the perceived fact *a policeman knocks on the door of the office at 5:53 p.m* (indeed *Mary knocks on the door at 6 p.m.* is not inconsistent with *a policeman knocks on the door at 5:53 p.m*). Thus there is not *mismatch-based surprise*. But Bill gets very *astonished* by perceiving the fact *a policeman knocks on the door of the office at 5:53 p.m.*. Indeed Bill retrieves the information concerning *a policeman knocking on the door of the office at 5:53 p.m.* from his background knowledge and recognizes the implausibility of the perceived fact given what he knows ("I wouldn't have expected to perceive a policeman knocking on the door of my office!").

Let us consider more carefully the two notions of *mismatch-based surprise* and *astonishment* from a qualitative and quantitative point of view. We want to specify the mental configurations associated with these two emotional responses and to provide the criteria to quantify them (to measure their intensity).

Definition 1: Mismatch-based Surprise (given the conflict between a perceived fact and a scrutinized representation). The cognitive configuration of *mismatch-based surprise* relative to the mismatch between a perceptual datum ψ and a scrutinized representation φ is defined by the following facts:

1. ψ is the agent's perceptual datum;
2. φ is the representation scrutinized by the agent and
3. the agent believes that φ and ψ are incompatible facts.

Formally:
$MismatchSurprise(\psi, \varphi) =_{def} Datum(\psi) \wedge Test(\varphi) \wedge Bel(\psi \rightarrow \neg\varphi)$.

Definition 2: (Retrieval-based) Astonishment. The cognitive configuration of *(retrieval-based) Astonishment* relative to a perceptual datum ψ is defined by the following facts:

1. ψ is the agent's perceptual datum;
2. the agent can retrieve from its background knowledge either the expectation $\neg\psi$ or the expectation that ψ that is, either the expectation that $\neg\psi$ or the expectation that ψ is "mentally" available at a background level.

Formally:
$Astonishment(\psi) =_{def} Datum(\psi) \wedge (Background(\neg\psi) \vee Background(\psi))$[11]

Definition 3: Intensity of Mismatch-based Surprise (given the conflict between a perceived fact and a scrutinized representation). The *mismatch-based surprise* relative to the mismatch between a perceptual datum ψ and a scrutinized representation φ has intensity equal to (or higher than) *c* if and only if the probability assigned to the scrutinized expectation that φ (invalidated by the perceived fact ψ) is equal to (or higher than) *c*.
Formally:
$IntensityMismatchSurprise(\psi, \varphi) \geq c =_{def} MismatchSurprise(\psi, \varphi) \wedge P(\varphi) \geq c$
$IntensityMismatchSurprise(\psi, \varphi) > c =_{def} MismatchSurprise(\psi, \varphi) \wedge P(\varphi) > c$
$IntensityMismatchSurprise(\psi, \varphi) = c =_{def} MismatchSurprise(\psi, \varphi) \wedge P(\varphi) = c$.

Definition 4: Intensity of (Retrieval-based) Astonishment. The *(retrieval-based) astonishment* relative to a perceptual datum ψ has intensity equal to (or higher than) *c* if and only if the probability assigned to $\neg\psi$ (the negation of the perceived fact) is equal to (or higher than) *c*.
Formally:
$IntensityAstonishment(\psi) \geq c =_{def} Astonishment(\psi) \wedge P(\neg\psi) \geq c$
$IntensityAstonishment(\psi) > c =_{def} Astonishment(\psi) \wedge P(\neg\psi) > c$
$IntensityAstonishment(\psi) = c =_{def} Astonishment(\psi) \wedge P(\neg\psi) = c$.

According to definitions 3 and 4 the intensity of *(retrieval-based) astonishment* is equal to the probability assigned to the opposite of the perceived fact (we can call it *degree of unexpectedness* of the perceived fact as in Ortony and Partridge, 1987) whereas the intensity of *mismatch-based surprise* is equal to the probability assigned to the formula invalidated by the perceived fact.
Let us discuss some formal properties of *(retrieval-based) astonishment* and *mismatch-based surprise*.

[11] Note that this definition is equivalent to the following definition: $Astonishment(\psi) =_{def} Datum(\psi) \wedge (\langle retrieve(\neg\psi)\rangle \top \vee \langle retrieve(\psi)\rangle \top)$.

Proposition 1.[12]

$\models_{Surp} IntensityMismatchSurprise(\psi, \varphi) = c \wedge (Background(\psi) \vee Background(\neg\psi))$
$\rightarrow IntensityAstonishment(\psi) \geq c$

The previous proposition says that if the agent is surprised by the mismatch between the perceptual datum ψ and a scrutinized expectation that φ and this surprise has intensity *c* then if the agent has either a background available expectation that ψ or a background available expectation that $\neg\psi$ then the intensity of astonishment is higher than *c*. Therefore *(retrieval-based) astonishment*s are by nature more intense than *mismatch-based surprises*. The reader should also note that the two dimensions of surprise are not necessarily complementary (the sum of the two is not necessarily equal to *1*). Indeed I could be surprised with intensity *0.5* by the mismatch between the perceptual datum ψ and the scrutinized expectation that φ and be astonished with intensity *0.7* by the recognized implausibility of ψ. Thus the two kinds of surprise are both qualitatively and quantitatively different.
Often *mismatch-based surprise* and *(retrieval-based) astonishment* occur together after having perceived a certain fact ψ. According to proposition 1 the intensity of (retrieval-based) astonishment is higher than the intensity of mismatch-based surprise. Consider next scenario.

Example 3. Imagine a person walking along the Thames. The person is scrutinizing whether *there is the tower of London* (φ) and is attributing a high probability to this fact. Suddenly the person turns the eyes toward the river and perceives *there is a whale* (ψ) (see the recent facts in London). The person gets *surprised* because of the recognition of the incompatibility between ψ and φ. Indeed the person believes that $\psi \rightarrow \neg\varphi$. But he also gets highly *astonished*. Indeed the person recognizes (after having retrieved from his background knowledge the information about ψ) the implausibility of the fact *there is a whale* (or even stronger the impossibility of the fact *there is a whale*). The intensity of the *astonishment* is equal to the probability assigned to $\neg\psi$.

3.2 Inference-based astonishment

In the previous paragraph we have defined astonishment the kind of surprise which involves a recognized implausibility of a perceived fact φ. We have assumed that the recognition of implausibility of the perceived fact φ is based on the mental availability of either the expectation that φ or the expectation that $\neg\varphi$. Indeed according to definition 2 retrieval-based astonishment concerns those background passive expectations that the agent can retrieve from the background level. As noticed by Ortony & Partridge (Ortony and Partridge, 1987) surprise can also arise from an inconsistency between an *implicit passive expectation* and the input proposition. With *implicit expectations* they mean all those facts that can be inferred from the explicit beliefs by few and simple deductions (see Figure 1).

[12] Formal proofs of theorems, lemmas and propositions are given in an extended version of this paper (Lorini and Castelfranchi, 2006a).

We think that Ortony & Partridge's distinction is relevant for a model of surprise and that in order to implement it formally we should relax the assumption of logical omniscience of the agent. In order to do it formally we should identify in the complete set of beliefs a subset of this and call it the set of *explicit beliefs* (or *belief base* as in the tradition of *Belief Revision*[13]). This is the set of beliefs that the agent can use to make inferences and which is not closed under classical inference[14]. Given the set of explicit beliefs we could define *implicit (passive) beliefs* as all those beliefs which can be inferred from the elements of the *belief base* (and which are not members of the belief base).
Having defined a set of *Explicit Beliefs* and a set of *Implicit Beliefs,* we can make more precise our definition of *Astonishment*. Indeed we can account for the *astonishment* due to a recognized conflict between a *post-hoc belief or expectation* (a belief which is inferred from the explicit beliefs and which was implicit before the perception) and the incoming input data: we call it *Inference-based Astonishment*.
Since the distinction between explicit and implicit belief is not formally specified under the present analysis we only give here a verbal characterization of *Inference-based Astonishment*.

Definition 5: (Inference-based) Astonishment. The cognitive configuration of *(Inference-based) Astonishment* relative to a perceptual datum ψ is defined by the following facts:

1. ψ is the agent's perceptual datum (something perceived by the agent);
2. the agent can infer and effectively infer $\neg\psi$ from its explicit beliefs (when $\neg\psi$ was the content of an implicit belief before the perception).

We should also consider as a matter of completeness all those cases of *post-hoc reconstruction of the probability* of the perceived event. This would allow us to generalize definition 5. In those cases while attempting to assimilate/integrate the perceived datum the agent "derives" that the event is not so probable (this is different from inferring some fact which is incompatible with the perceived fact). While asking to himself: was this unpredicted event/datum predictable? it reconstructs the probability of the event and conclude that "I would never had expected that". Therefore the intensity of *inference-based astonishment* relative to the perceived fact ψ must depend on the probability assigned to $\neg\psi$ (the higher the probability assigned to $\neg\psi$, the more intense the astonishment).

We have provided two different notions of *astonishment.* On one side (definition 2) after perceiving ψ there is a simple *retrieval* of either the expectation that ψ or the expectation that $\neg\psi$ when either the expectation that ψ or the expectation that $\neg\psi$ is mentally available at the background (passive) level. On the other side (definition 5) either the negation of the perceived fact is *inferred* from the explicit beliefs or there

[13] See for instance Hansson (1999) for a complete account of belief revision applied to belief bases.
[14] Obviously we assume that the *expectation under scrutiny* is a special kind of *explicit belief* (see Figure 1).

is a post-hoc reconstruction of the probability of the perceived fact (a probabilistic inference). In both cases some *mental operation* must be done in order to make the agent aware of the implausibility of the perceived fact.
We conclude this section by summarizing our basic ontology of *on-line* surprise (whose cognitive configuration is obtained during the perceptual phase and before an eventual belief reconsideration). In our view at least three species must be considered: *surprise based on the mismatch between a representation under scrutiny and an incoming input* (definition 1)*, retrieval-based astonishment* (definition 2) and *inference-based astonishment* (definition 5).

3.3 Some comments

Let us stress more in this section the main differences between our approach and Ortony & Partridge's approach by making explicit the main important issues that are neglected in their model and that our model tries to clarify.
Ortony & Partridge's model does not capture in our view the important distinction between the previous two kinds of astonishments (retrieval-based astonishment and inference-based astonishment). Ortony & Partridge's model is only focused on inference-based astonishment and completely neglects to account for the other important kind.
In *inference-based astonishment*, the subject did not in fact derive the prediction/expectation that $\neg\varphi$ before perceiving φ (the prediction is just potential and implicit in its mind). While attempting to assimilate/integrate the new data he infers from his explicit beliefs the opposite. Therefore the mental operation involved in this kind of astonishment is an *inferential* action[15]: it transforms some potential and implicit expectation (or belief) into an explicit and scrutinized one. This is exactly the content of the informal definition 5 given in section 3.2.
In *retrieval-based astonishment* on the contrary, when perceiving φ a pre-existent expectation that φ (or a pre-existent expectation that $\neg\varphi$) is available (it can be retrieved from the background level even without a constructive inferential process). Indeed in our view an agent has always a certain number of accessible beliefs and expectations in background (at an unconscious and automatic level) and these expectations and beliefs in background must be distinguished from the representation under scrutiny formally identified as a *test formula* (see Figure 1 and section 2.3). When perceiving φ, retrieval-based astonishment may simply arise from the automatic retrieval of either the background probabilistic expectation that φ or the background probabilistic expectation that $\neg\varphi$. Therefore the mental operation involved in retrieval-based astonishment is a *retrieve mental operation* which transforms some background expectation (or belief) into a scrutinized expectation. This is exactly the content of the formal definition 2 given in section 3.1.
In our view Ortony & Partridge's model does not capture this distinction between 1) surprise arising from the recognition of implausibility of the perceived fact due to an inferential process from my explicit beliefs and 2)
surprise arising from the recognition of implausibility of the perceived fact due to a

[15] As noticed in the previous section it could also be a probabilistic inference (a post-hoc reconstruction of the probability of the perceived fact).

retrieval of a background expectation. The incompleteness of Ortony & Partridge's model is due to the lack of distinction between *background expectations and representations* on one side and *implicit expectations and beliefs* on the other side (indeed they only account for the second kind). This distinction is relevant in our approach and it gives us the possibility to articulate a richer typology of surprise.

Moreover, in our model there are two parallel components and paths for surprise, and there are two parameters that we should take into account in order to quantify surprise (see figure 2 below).
(i) I can have an expectation under scrutiny whose content is φ (the expected event or entity): when this prediction is invalidated, happens to be wrong, this means that I perceive something different. In other word there is an input datum ψ mismatching with φ. Even nothing is something: also the absence of any object when I was expecting and scrutinizing φ, that is the fact that φ does not happens ($\neg\varphi$) is in any case an unpredicted/unexpected input datum which invalidates the representation under scrutiny φ.
(ii) Having perceived ψ, the expectation that ψ (or the expectation that $\neg\psi$) is available at the background and unconscious level (or the expectation that $\neg\psi$ is inferred from explicit beliefs and expectations).
We claim that, on the one side, surprise is a function of the probability of the invalidated fact under scrutiny (φ); while on the other side it is a function of the probability of the perceived fact ψ . On the one side the more certain was my scrutinized expectation, the more probable is φ, the more surprise I am[16] (see definition 3 in section 3.1). On the other side the more unpredictable, the more unexpected ψ (the more expected $\neg\psi$), the more astonished I am (see definition 4 in section 3.1 as well as the generalization of definition 5 which deals with probabilistic inference). To distinguish these two facets, components, and processes we have proposed to use for the former case the term *mismatch-based surprise* (the signal of the invalidation of the expectation under scrutiny), and the term *astonishment* (either retrieval-based astonishment or inference-based astonishment) for the latter case. Ortony & Partridge seem to consider surprise only the second phenomenon and path. Indeed according to their model intensity of surprise only depends on the degree of "unexpectedness" of the perceived fact [17].
But not always the surprise processing involves the two paths. Indeed one can be surprised by some perceived fact ψ one did not expect without having to expect and test something else which is evaluated to be incompatible with ψ: not necessarily the *astonishment* presupposes the *mismatch-based surprise*. Moreover one can be surprised by some perceived fact ψ which is evaluated to be incompatible with some scrutinized fact φ without having to be astonished by the recognized implausibility of ψ: not necessarily a *mismatch-based surprise* entails an *astonishment* as a felt reaction.

[16] At a meta-level too we might say that the mismatch was unexpected.

[17] The same criticism can be addressed toward all those computational models which claim that surprise is simply a function of *unexpectedness* of the incoming input and which neglect the dimension *strength of the invalidated expectation* (Macedo and Cardoso, 2001; Meyer et al., 1997; Ortony and Partridge, 1987). Other models based on Information theory claim that Surprise is a function of the *distance* between prior probabilities and posterior probabilities after the conditioning on the set of perceived data (see Baldi, 2004 for instance). For the same reasons we believe that this last approach is incomplete since it is unable to provide qualitative distinctions inside the surprise phenomenon.

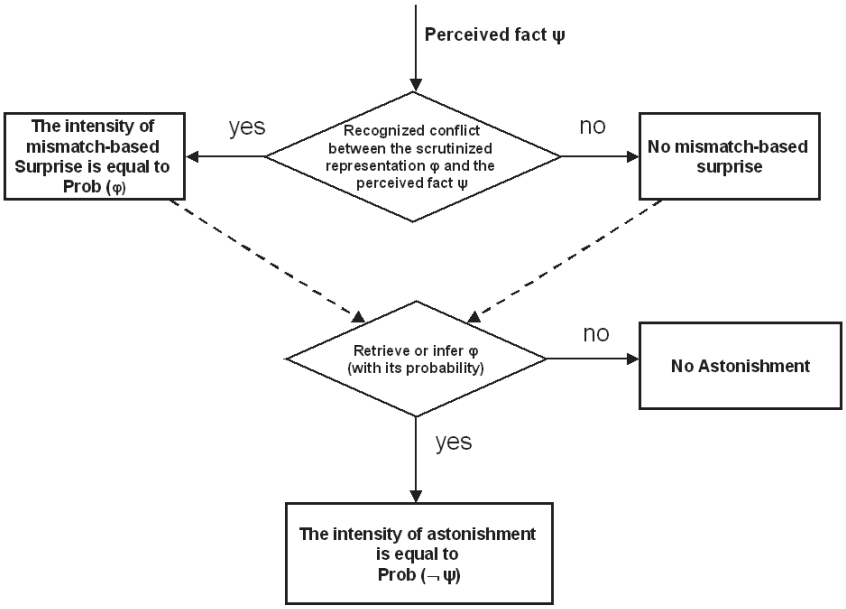

Figure 2: Surprise processing

4 Surprise and belief change

As some psychologists have stressed (Meyer et al., 1991, 1997) surprise can culminate in a process of belief change. The aim of the following analysis is to suggest some interesting ways a formal model of *cognitive surprise* can be integrated with a formal model of *cognitive belief change*.

Formal approaches to belief revision are mainly interested in finding rationality principles and postulates driving belief change (this is for instance the main purpose of the classical AGM theory Alchourron et al., 1985). All those models implicitly assume that when the agent perceives some fact φ the perception is always a precursor of a belief change with φ. Thus the main problem with AGM theory is a missed identification of the precursors of belief change.

Our attempt here is to clarify *under what conditions* belief revision *should be* triggered after having perceived a certain fact. We claim that surprise plays a crucial role in triggering this kind of process and that it is implausible to assume that realistic cognitive agents revise their beliefs with φ every time they perceive a new fact φ. Realistic and non-omniscient cognitive agents are situated in complex environments where many tasks must be solved. Since accurate belief revision and update require time and considerable computational costs, realistic cognitive agents need some mechanism which is responsible:

1) for signaling the global inconsistency of the knowledge base with respect to the incoming input and

2) for the revision of beliefs and expectations of the agent.

One of the adaptive functions of surprise is exactly this.

Belief change in cognitive agents is triggered by very surprising incoming input. The intensity of surprise relative to the incoming input "signals" to the agent that things are not going as expected and that the knowledge of the environment must be reconsidered. Indeed wrong beliefs generally lead to bad performances and to failure in the intention and goal fulfillment.

On the other hand resource bounded cognitive agents do not generally reconsider their beliefs and expectations when the input data are not recognized to be incompatible or implausible with respect to their pre-existent knowledge. Indeed it is not convenient for the survival of the agent to update or reconsider beliefs every time a new fact is perceived. When the world flows as expected and we are not aware of the inadequacy of our knowledge of the world, we do not need to criticize and reconsider this knowledge. Indeed reconsidering beliefs after every perception would strongly interfere with the agent's ongoing performance and would continuously divert its attention away from its intentionally driven activity.

A model of cognitive belief change should be able to account for this trade-off between *extensive belief change* triggered by surprise and *belief change avoidance* when perception does not generate surprise.

4.1 Dealing with unexecutable updates

In Kooi (2003) a combination of the dynamic epistemic logic of Gerbrandy (Gerbrandy, 1999; Gerbrandy and Groeneveld, 1997) with the probabilistic logic of Fagin &

Halpern is given. This combination results in a probabilistic dynamic epistemic logic where it is possible to talk about beliefs and probabilities as well as information change for beliefs and probabilities. In this probabilistic extension of Gerbrandy's logic of information update the symbol $\varphi!$ is introduced. $\varphi!$ is the process of updating beliefs with an arbitrary sentence φ.[18]
The aim of this section is to suggest a way to modify the framework given in Kooi (2003) and Gerbrandy (1999); Gerbrandy and Groeneveld (1997) in order to investigate the role of surprise in information change.
In order to do this we must import *update processes* into our formal language $\mathcal{L}_{\mathcal{SURP}}$. UPD is the set of *update processes* defined as the smallest set such that:

- if $\varphi \in PROP$ then $\varphi! \in UPD$.

We call $\mathcal{L}_{\mathcal{SURP}+}$ the new extended language with *update processes*. The new language $\mathcal{L}_{\mathcal{SURP}+}$ is given by the following rule in extended Backus-Naur Form:

$$\Phi ::= p|\neg\Phi|\Phi_1 \wedge \Phi_2|Bel\Phi|\,[\alpha]\,\Phi|d_1P(\Phi_1) + ... + d_nP(\Phi_n) \geq c|Test(\varphi)|Datum(\varphi)|\,[\varphi!]\,\Phi$$

where $p \in \Pi$, $\varphi \in PROP$, $\alpha \in ACT$, $d_1, ..., d_n, c$ are real numbers and $\varphi! \in UPD$. Semantics of formulas in $\mathcal{L}_{\mathcal{SURP}+}$ is the same semantics given for formulas in $\mathcal{L}_{\mathcal{SURP}}$ (see section 2.2). We only need to provide semantics for formulas of the form $[\varphi!]\,\Phi$. This is given next.[19]

- $M, w \vDash [\varphi!]\,\Phi \Longleftrightarrow \forall(M', w')$ if $(M', w') \in R^{\varphi!}(M, w)$ then $M', w' \vDash \Phi$

We suppose that $R^{\varphi!}(M, w)$ is defined according to the following definition 6.

Definition 6. Given a model $M =< W, B, R_0, R_1, R_2, P, TEST, DATA, \pi >$, a world $w \in W$ and a propositional formula $\varphi \in PROP$ we suppose that

$$\text{EITHER } R^{\varphi!}(M, w) = \emptyset \text{ OR } R^{\varphi!}(M, w) = (\widetilde{M^{\varphi}}, \widetilde{w^{\varphi}}).$$

Moreover we suppose that $\widetilde{M^{\varphi}}$ and $\widetilde{w^{\varphi}}$ are defined as follows.

1. $\widetilde{w^{\varphi}} = w$.
2. $\widetilde{M^{\varphi}} =< W, \widetilde{B^{\varphi}}, R_0, R_1, R_2, \widetilde{P^{\varphi}}, TEST, DATA, \pi >$ where $\widetilde{P^{\varphi}}(w) = (\widetilde{W_w^{\varphi}}, \widetilde{X_w^{\varphi}})$ and $\widetilde{W_w^{\varphi}}$, $\widetilde{X_w^{\varphi}}$, $\widetilde{B^{\varphi}}$ are defined according to the following (a), (b) and (c).

[18] Richer logics of information update have been proposed. In Baltag et al.'s logic of information update (Baltag et al., 1998) for instance complex communicative actions are described in terms of *action models*, which stand for complex events that carry information for agents. Different kinds of informational scenarios in a multi-agent setting can be described in this logic. For instance we can describe scenarios where not all agents have the same observational access to what is happening in reality. In van Benthem et al. (2006) probabilities are added to Baltag et al.'s framework in order to reason about probabilistic information in a multi-agent setting and to describe how belief and probability update is affected by the reliability of the source of information.

[19] We generalise here the standard approach of dynamic logic where actions are interpreted as transitions between worlds in a Model (see also Van Linder et al., 1997; Meyer et al., 1999). Indeed we interpret update processes as transitions between pairs (Model, world).

(a) for all $w \in W$: $\widetilde{B^{\varphi}}(w) = \{w' | w' \in B(w)\ and\ M, w' \models \varphi\}$.

(b) for all $w \in W$: $\widetilde{W_w^{\varphi}} = \{w' \in W_w | M, w' \models \varphi\}$.

(c) for all $w \in W$ and $w' \in \widetilde{W_w^{\varphi}}$: $\widetilde{X_w^{\varphi}}(w') = \frac{X_w(w')}{X_w(||\varphi||^{W_w})}$.

According to definition 6 updating with φ either cannot be performed or yields an updated model $\widetilde{M^{\varphi}}$ which differs from the original model only with respect to the accessibility relations for *Bel* modal operator and the probability functions. When an update with φ is *successfully* performed the original model M is transformed into the updated model $\widetilde{M^{\varphi}}$ in such a way that for all worlds *w* all alternatives that an agent considers possible where φ does not hold are removed and worlds where φ does not hold are removed from the sample space W_w. Moreover for all worlds *w* the probability function is redefined according to Condition 2(c).
In Kooi (2003); Gerbrandy and Groeneveld (1997); Gerbrandy (1999) it is assumed that an update with φ is always executable i.e. the authors suppose that $R^{\varphi!}(M, w)$ is never empty and always yields the updated model $\widetilde{M^{\varphi}}$. Thus these theories of information update assume that the formula $\langle\varphi!\rangle \top$ is valid. Here we suppose that belief update is triggered only under certain specific preconditions. This implies that in our view $R^{\varphi!}(M, w)$ may be empty (see definition 6). This is the most striking difference between our version of belief update and standard versions of it. We will try to characterize the necessary preconditions for belief update in the next section and to investigate the role of surprise in the process.

The first relevant aspect to verify is whether the previous model transformation guarantees that the semantic constraints given in section 2.3 are preserved. This is indeed the case.

Lemma 1.

If M is a *SURPRISE model* then $\widetilde{M^{\varphi}}$ is a *SURPRISE model* too.

Let us suppose in a way similar to Kooi (2003) that an update with φ can be performed only if the agent does not assign zero probability to φ. This assumption is made explicit in our framework by the next postulate:

$(NotZero_{Upd})\ \langle\varphi!\rangle \top \rightarrow P(\varphi) > 0.$

This property corresponds to the following semantic constraint:

- for every $w \in W$ if $R^{\varphi!}(M, w) \neq 0$ then $X_w(||\varphi||^{W_w}) > 0$.[20]

You should notice that under this requirement formula $Bel\varphi \rightarrow [\neg\varphi!] \perp$ becomes valid, that is if an agent believes that φ holds then an update with φ cannot be executed. This

[20] This assumption is made in order to prevent from dividing by zero (condition 2c in definition 6) when redefining the probability function after an update with a sentence with probability zero. More general approaches to updating with sentences with probability zero are discussed in Halpern (2001, 2003).

implies that under the present framework *belief revision* with inconsistent information is left unspecified.[21]
We also postulate that an agent has always epistemic access to all executable updates that is, if an update with sentence φ is executable then the agent believes that the update with φ is executable. Formally:

$(Access_{Upd})\ \langle\varphi!\rangle\top \rightarrow Bel\,\langle\varphi!\rangle\top.$

This property corresponds to the following semantic constraint:

- for every $w \in W$ if there is a w' such that $w' \in B(w)$ and $R^{\varphi!}(M, w') = 0$ then $R^{\varphi!}(M, w) = 0$.

Before starting to investigate some formal consequences of our definition of belief update we provide the following definition of *objective formula*.

Definition 7. We define the set of *objective formulas* $OBJ = \{o_1, o_2, ...\}$ as the smallest set such that:

- if $\varphi \in PROP$ then $\varphi \in OBJ$ (propositional formulas are objective formulas);
- if $\varphi \in PROP$ then $Test(\varphi)$ and $\neg Test(\varphi) \in OBJ$ (test formulas and negation of test formulas are objective formulas);
- if $\varphi \in PROP$ then $Datum(\varphi)$ and $\neg Datum(\varphi) \in OBJ$ (perception formulas and negations of perception formulas are objective formulas);
- if $o_1 \in OBJ$ and $\alpha \in ACT$ then $[\alpha]\,o_1$ and $\langle\alpha\rangle\,o_1 \in OBJ$.

We can now prove that the principles summarized in the following Lemma 2 are sound given the semantics of update processes (definition 6).

Lemma 2.

$(Upd_1)\ \langle\varphi!\rangle\,\Phi \rightarrow [\varphi!]\,\Phi$
$(Upd_2)\ o_1 \rightarrow [\varphi!]\,o_1$ where o_1 is an objective formula
$(Upd_3)\ Bel(\varphi \rightarrow \langle\varphi!\rangle\,\Phi) \rightarrow [\varphi!]\,Bel\Phi$
$(Upd_4)\ \langle\varphi!\rangle\,Bel\Phi \rightarrow Bel(\varphi \rightarrow [\varphi!]\,\Phi)$
$(Upd_5)\ (\sum_{i=1}^{n} d_i P(\varphi \wedge \langle\varphi!\rangle\,\Phi_i) \geq cP(\varphi)) \rightarrow [\varphi!]\,(\sum_{i=1}^{n} d_i P(\Phi_i) \geq c)$
$(Upd_6)\ (\langle\varphi!\rangle\,(\sum_{i=1}^{n} d_i P(\Phi_i) \geq c) \rightarrow (\sum_{i=1}^{n} d_i P(\varphi \wedge [\varphi!]\,\Phi_i) \geq cP(\varphi))$
$(Upd_7)\ [\alpha]\,\langle\varphi!\rangle\,\Phi \rightarrow [\varphi!]\,[\alpha]\,\Phi$

(Upd_1) establishes that belief updates are deterministic. According to (Upd_2) the truth value of an objective formula does not change after a belief update. (Upd_3), (Upd_4), (Upd_5) and (Upd_6) describe how beliefs and probabilities change after an update. According to (Upd_7) the effects of an update process on a model are independent from

[21] For *belief revision* with inconsistent information see for instance Herzig and Longin (2002); van Benthem (2006).

the fact that the update process may be executed after or before a sequence of actions (a sequence where each element is either an *atomic action* or a *perceptual action* or a *retrieve mental operation*).

Finally we can precisely define our extended logic of surprise with update processes. We call *SURPRISE+* the logic axiomatized by the axioms and inference rules of the logic *SURPRISE* (see section 2.3) plus the previous nine principles for update processes (Upd_1)-(Upd_7), $(Access_{Upd})$ and $(NotZero_{Upd})$. Moreover we write $\vdash_{Surp+} \varphi$ if φ is a theorem of *SURPRISE+*.

We are able to prove by the seven principles summarized in Lemma 2 and the previous postulates $Access_{Upd}$ and $Incl_{Bel/Prob}$ (section 2.3) that two compact reduction principles for beliefs and updates on side and probabilities and updates on the other side follow from the axiomatic system of our logic. These two principles are similar to Gerbrandy's reduction principle for beliefs and updates (Gerbrandy and Groeneveld, 1997; Gerbrandy, 1999) and Kooi's reduction principle for probabilities and updates (Kooi, 2003). These results are summarized in the following theorem.

Theorem 1.

$(Upd_8) \vdash_{Surp+} Bel(\varphi \rightarrow [\varphi!]\,\Phi) \vee [\varphi!] \perp \longleftrightarrow [\varphi!]\,Bel\Phi$
$(Upd_9) \vdash_{Surp+} (\sum_{i=1}^{n} d_i P(\varphi \wedge [\varphi!]\,\Phi_i) \geq cP(\varphi)) \vee [\varphi!] \perp \longleftrightarrow [\varphi!]\,(\sum_{i=1}^{n} d_i P(\Phi_i) \geq c)$

Several interesting properties of update processes follow from theorem 1 and the principles given in Lemma 2. Let us consider only some of them.

Proposition 2.

$(Upd_{10}) \vdash_{Surp+} [\varphi!]\,Bel\varphi$
$(Upd_{11}) \vdash_{Surp+} Bel^m o_1 \rightarrow [\varphi!]\,Bel^m o_1$ for each $m > 0$
$(Upd_{12}) \vdash_{Surp+} P(\varphi|\psi) = c \rightarrow [\psi!]\,P(\varphi) = c$ where $P(\varphi|\psi) = \frac{P(\varphi \wedge \psi)}{P(\psi)}$

According to (Upd_{10}) after an update with φ the agent believes that φ holds. According to (Upd_{11}) for each *m*-level nested belief that o_1 holds (where o_1 is an objective formula), the *m*-level nested belief is preserved after a belief update. (Upd_{12}) shows the strong similarity between updating with propositional formulas in our framework and classical bayesian updating.[22]

4.2 Surprise-based belief update

We have noticed in the previous section that a relevant difference exists between the present approach to belief update and some standard approaches (Kooi, 2003; Gerbrandy and Groeneveld, 1997; Gerbrandy, 1999). Differently from standard approaches

[22] The same result is obtained in Kooi (2003).

we have supposed that belief update is triggered only under certain specific preconditions and is not always executable. The aim of this section is characterize some of those necessary preconditions for belief update and to show that surprise plays a crucial role in triggering this mental process.
We begin with the formalization of our general intuition by supposing that two necessary preconditions for belief update are expressed by the following two additional principles $(NecTrig1)$ and $(NecTrig2)$.

$(NecTrig1)$ $\langle\varphi!\rangle\top \rightarrow Datum(\varphi)$[23]
$(NecTrig2)$ $\langle\varphi!\rangle\top \wedge Test(\psi) \rightarrow Bel(\varphi \rightarrow \neg\psi) \vee Background(\varphi) \vee Background(\neg\varphi)$[24]

According to principle $(NecTrig1)$ an agent cannot update its beliefs with sentence φ unless φ is something that the agent has perceived (φ is a perceptual datum of the agent). According to principle $(NecTrig2)$ if the agent is focused on the expectation that ψ then the agent cannot update is beliefs with φ unless either the agent recognizes a contradiction between φ and its scrutinized expectation that ψ or φ (or $\neg\varphi$) is the content of an available background expectation. Both principles formally express the following postulate.
An agent can reconsider its previous knowledge with some piece of information φ only if:
1) φ is some piece information that agent has perceived and which is collected as a perceptual datum and
2) either the agent recognizes (is aware of) the contradiction and incompatibility between the perceptual datum φ (the object of its perception) and its scrutinized expectation or the probabilistic expectation that φ (or the expectation that $\neg\varphi$) is (mentally) available at a background level.
Thus according to the previous postulate if an agent is not aware of the inconsistency between the perceptual datum φ and its actual scrutinized expectation that ψ and does not have access to the information concerning the plausibility of φ then the agent cannot revise its knowledge base on the basis of the perceptual datum.
The following example is given in order to defend the plausibility of the present postulate.

Example 4. Mary goes shopping downtown. She is looking for a nice pair of shoes for New Year's Eve party. She remembers having heart from Bill that a well-stocked shoe shop has been opened at the main square of the town. Mary trusts Bill since she thinks that Bill gives always good advice. Thus she decides to reach the main square of the town in order to find the shoe shop. Now Mary expects that φ_1 = *she will*

[23] This postulate corresponds to the following semantic constraint: for every $w \in W$ if $R^{\varphi!}(M, w) \neq \emptyset$ then $\varphi = DATA(w)$.

[24] Given the definition $Background(\varphi) =_{def} \langle retrieve(\varphi)\rangle\top$ and property $(Perc_1)$ of perceptual actions (see section 2.3) we can express the semantics corresponding to this principle by the following first order formula:

for every $w \in W$ if $R^{\varphi!}(M, w) \neq \emptyset$ and $TEST(w) = \psi$ and there is a w' such that $w' \in B(w)$ and $R_1^{observe(\varphi)}(w') \neq \emptyset$ and $R_1^{observe(\psi)}(w') \neq \emptyset$ then $R_2^{retrieve(\varphi)}(w') \neq \emptyset$ or $R_1^{retrieve(\neg\varphi)}(w') \neq \emptyset$.

find a shop selling a nice pair of shoes at the main square with high probability and focuses her attention on this. When walking toward the shop Mary observes $\varphi_2 =$ *there is a Japanese restaurant in the corner of the street*. Nevertheless Mary does not care about φ_2. Indeed: 1) φ_2 is not evaluated to be incompatible with φ_1 and 2) Mary does not have a background available expectation that φ_2 nor a background available expectation that $\neg\varphi_2$ which makes her able to recognize the implausibility of the perceived fact φ_2. Thus Mary does not reconsider her knowledge base according to what she has perceived since both a recognition of implausibility of the perceived fact and a recognition of incompatibility between the perceived fact and the scrutinized expectation that φ_1 are lacking.
Finally Mary arrives at the main square of the town where she expects to find the shoe shop and to buy a nice pair of shoes. But Mary sees that no shop is there. Mary recognizes the inconsistency between her scrutinized expectation (φ_1) and what is being perceived. Indeed there is not a shoe shop at the place where she expected to find a shoe shop. Since Mary is aware of the incompatibility between the perceived fact and her actual scrutinized expectation she can reconsider her belief base according to the perceptual datum.

Given the definitions of astonishment and mismatch-based surprise (section 3.1) and supposing that the previous two principles $(NecTrig1)$ and $(NecTrig2)$ are added to our logic *SURPRISE+* the following becomes a provable theorem.

Proposition 3.

$$\vdash_{Surp+} \langle\varphi!\rangle \top \wedge Test(\psi) \rightarrow MismatchSurprise(\varphi, \psi) \vee Astonishment(\varphi)$$

According to Proposition 3 if the agent is focused on the expectation that ψ then the agent cannot revise its knowledge base with the perceived fact φ unless either the agent gets surprised by the mismatch between the perceptual datum φ and the scrutinized expectation that ψ or the agent gets astonished by the recognized implausibility of φ. This proposition expresses a general cognitive principle: belief update with a perceived fact φ is triggered only if the agent is surprised or astonished by the perception of φ, that is
Some form of surprise is a necessary precondition for belief update.
This is for us a crucial principle for designing resource bounded cognitive agents which are focused on a small portion of their complete informational state and which need some mechanism for "signaling" that beliefs must be updated.

After having characterized two "necessary" preconditions for triggering belief update we move toward a brief investigation of the "necessary and sufficient conditions". We only provide here some general intuitions about this issue.
It has been noticed by some psychologists (Meyer et al., 1997) that the trigger of a belief update process depends on the intensity of surprise associated with the perception of some fact φ: the higher the intensity of surprise relative to the perception of φ, the higher the probability that the agent will revise its knowledge with ψ.

In our view a first rough approximation of the necessary and sufficient preconditions for belief update is obtained by introducing the previous dimension: the intensity of surprise associated with the perception of φ.
We suggest the following as a plausible solution to the identification of the "necessary and sufficient" preconditions for belief update.
We establish that if the agent is scrutinizing the expectation that φ then it updates its belief base with ψ if and only if:

- ψ is a perceptual datum and
- either the agent gets surprised by the mismatch between the perceptual datum ψ and the scrutinized expectation that φ and the intensity of mismatch-based surprise exceeds a given threshold Δ or
- the agent gets astonished by the recognition of implausibility of ψ and the intensity of astonishment exceeds threshold Δ.

We can express formally the previous principle.

$(NecSuffTrig)\ \langle\varphi!\rangle \top \wedge Test(\psi) \longleftrightarrow Test(\psi) \wedge (IntensityMismatchSurprise(\varphi, \psi) > \Delta \vee IntensityAstonishment(\varphi) > \Delta)$

Let us note two relevant facts. On one hand we want to emphasize that both personality factors and motivational factors can affect the value of the threshold Δ and that the value of Δ changes due to the evolution and dynamics of goals and intentions. Since Δ has an intrinsic dynamic nature, its value is not in principle the same for all possible worlds *w* in a model. Nevertheless it seems plausible to state that the higher is the *motivational relevance* of the perceived fact (more important is φ given actual goals and intentions of the agent) and the lower is the value of Δ. This implies that I am more prone to revise my knowledge when I perceive something which is *relevant* with respect to my actual motivations than when I perceive something which is completely *irrelevant* with respect to my actual motivations.
On the other hand we want to emphasize that the previous characterization $(NecSuffTrig)$ of "necessary and sufficient preconditions" for belief update is somehow still unsatisfactory. It must be stressed that a more articulated model of the process would require a distinction between *belief change* vs. *belief rejection*. Indeed after having been surprised by the perceived fact φ, not necessarily the agent "decides" to update its beliefs. The agent may simply decide to reject φ if the source of information is evaluated to be unreliable (Castelfranchi, 1997). This means that once the agent has been surprised, the possibility of updating beliefs with a perceived fact φ also depends on the reliability assigned to the source of information (reliability of the sensors or reliability of the communicative source etc...). Indeed after being surprised by the perception of φ, I am more prone to revise my knowledge with φ (instead of rejecting the perceptual datum φ) when I consider my sensors to be reliable ("so it is not a hallucination!") than when I consider my sensors to be unreliable.

5 Conclusion

We have provided in this paper a conceptual and formal clarification of the notion of *surprise* thanks to the elaboration of the ontology developed in section 3. Each kind of surprise has been associated with a particular phase of the cognitive processing and involves particular kinds of epistemic representations (representation and expectation under scrutiny, perceptual data, presupposed frame, background expectations and beliefs).

We have identified two main kinds of surprise: *mismatch-based surprise* and *astonishment*. The first has been defined as the surprise due to a recognized inconsistency between an expectation under scrutiny and a perceived fact. The second has been defined as the surprise due to the recognition of implausibility of the perceived fact where this recognition is based either on the retrieval of a background expectation or on some inferential process (classical deduction or probabilistic inference). We have compared our model with existing psychological models of surprise and shown that an analytic investigation of the concept is still missing and that in these models some important aspects of this cognitive phenomenon are ignored.

In the second part of the paper (section 4) we have investigated the role of surprise in triggering belief update. We think in fact that the notion of surprise should be exploited by current formal models of information update. We have provided several justifications of our theoretical position. Indeed on one hand we think that in designing cognitive agents we must relax the assumption that in principle any perception produces a reconsideration of pre-existent beliefs and expectations. Since realistic agents are by definition resource-bounded (Wasserman, 1999; Cherniak, 1986), they should not waste time and energy in reasoning out and reconsider their knowledge on the basis of every piece of information they get. To relax the previous assumption seems indeed a necessary desideratum to bridge the existing gap between formal models of belief change and cognitive theories of belief dynamics. On the other hand we think that after having relaxed the previous assumption we must look for the cognitive precursors of belief change. We have stressed that surprise is perhaps the most important causal precursor of belief change. We have presented a method to integrate surprise in a formal model of belief update and to investigate its functional role.

More work must be done in order to improve the present model. *From a strictly formal point view*. We have not yet completeness results for our modal logic of surprise. *From a more theoretical point view*. We have characterized several kinds of informational mental states such as scrutinized expectations (expectations on which the agent focuses its attention) and background expectations (expectations which are available at a mere automatic and effortless level). Moreover we have characterized mental processes which are responsible for modifying those scrutinized expectations (we have called them *retrieve mental operations*) by transforming one background expectation into a scrutinized one. We still miss a systematic explanation and formal account of why certain expectations rather other ones go in background and become accessible. Moreover we have not explained why certain expectations rather than other ones get scrutinized by the agent.

For the moment we leave unsolved these formal and theoretical problems and we postpone them to future work.

6 Acknowledgment

We are very grateful to Johan van Benthem and to the anonymous referees of this paper for their helpful comments on the content of our work.
Our research has been supported by the European Project "MindRACES: from Reactive to Anticipatory Cognitive Embodied Systems" (IST-511931).

References

Alchourron, C., Gardenfors, P., and Makinson, D. (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50:510–530.

Baldi, P. (2004). Surprise: A shortcut for attention? In Itti, L., Rees, G., and Tsotsos, J., editors, *Neurobiology of Attention*. Academic Press.

Baltag, A., Moss, L., and Solecki, S. (1998). The logic of public announcements, common knowledge and private suspicions. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*.

Bloch, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18:227–287.

Castelfranchi, C. (1997). Representation and integration of multiple knowledge sources: Issues and questions. In Cantoni, V., Di Gesù, V., Setti, A., and Tegolo, D., editors, *Human & Machine Perception: Information Fusion*, pages 235–254. Plenum Press, New York.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 3:200–219.

Cherniak, C. (1986). *Minimal rationality*. MIT Press, Cambridge.

Dretske, F. I. (1981). *Knowledge and the flow of information*. MIT Press, Cambridge.

Fagin, R. and Halpern, J. (1994). Reasoning about knowledge and probability. *Journal of the Association for Computing Machinery*, pages 340–367.

Fagin, R. and Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39–76.

Gerbrandy, J. (1999). *Bisimulations on Planet Kripke*. PhD thesis, University of Amsterdam, The Netherlands.

Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–196.

Halpern, J. Y. (2001). Lexicographic probability, conditional probability, and non-standard probability. In *Proceedings of Eight Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 17–30.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press, Cambridge.

Hansson, S. O. (1999). *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Kluwer, Dordrecht, Netherland.

Herzig, A. and Longin, D. (2002). Sensing and revision in a modal logic of belief and action. In *Proceedings of the fifteenth European Conference on Artificial Intelligence (ECAI02)*, pages 307–311.

Kahneman, D. and Miller, D. T. (1986). Norm theory: comparing reality to its alternatives. *Psychological Review*, 93:136–153.

Kahneman, D. and Tversky, A. (1982). Variants of uncertainty. *Cognition*, 11:143–157.

Kooi, B. P. (2003). Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information*, 12:381–408.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI84)*, pages 198–202.

Lorini, E. and Castelfranchi, C. (2006a). The cognitive structure of surprise: looking for basic principles. Technical report, http://www.istc.cnr.it/createhtml.php?nbr=83.

Lorini, E. and Castelfranchi, C. (2006b). The unexpected aspects of surprise. *International Journal of Pattern Recognition and Artificial Intelligence*, in press.

Macedo, L. and Cardoso, A. (2001). Modelling forms of surprise in an artificial agent. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*.

Matt, J., Leuthold, H., and Sommer, W. (1992). Differential effects of voluntary expectancies on reaction times and event-related potentials: evidence for automatic and controlled expectancies. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:810–822.

Meyer, J. J. C., van der Hoek, W., and van Linder, B. (1999). A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113(1-2):1–40.

Meyer, W.-U., Niepel, M., Rudolph, U., and Schutzwohl, A. (1991). An experimental analysis of surprise. *Cognition and Emotion*, 5:295–311.

Meyer, W. U., Reisenzein, R., and Schützwohl, A. (1997). Towards a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21:251–274.

Ortony, A. and Partridge, D. (1987). Surprisingness and expectation failure: Whats the difference? In *Proceedings of the tenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 106–108.

Paglieri, F. (2004). Data-oriented belief revision: Towards a unified theory of epistemic processing. In *Proceedings of STAIRS2004*.

Reisenzein, R. (2000). The subjective experience of surprise. In Bless, H. and Forgas, J., editors, *The message within: The role of subjective experience in social cognition and behavior*. PA: Psychology Press, Philadelphia.

Reisenzein, R., Meyer, W.-U., and Schutzwohl, A. (1996). Reactions to surprising events: A paradigm for emotion research. In *Proceedings of the 9th conference of the International Society for Research on Emotions*, pages 292–296, Toronto.

Rescher, N. (1976). *Plausible reasoning*. Van Gorcum, Assen.

Sommer, W., Leuthold, H., and Matt, J. (1998). The expectancies that govern the P300 amplitude are mostly automatic and unconscious. *Behavioral and Brain Sciences*, 21:149–150.

Tamminga, A. (2001). Expansion and contraction of finite states. *Studia Logica*, 68:1–16.

Tversky, A. and Koehler, D. J. (1994). Support theory: a nonextensional representation of subjective probability. *Psychological Review*, 101(4):547–567.

van Benthem, J. (2006). Dynamic logic for belief revision. *ILLC Publication*.

van Benthem, J., Gerbrandy, J., and Kooi, B. (2006). Dynamic update with probabilities. *ILLC Publication*.

Van Linder, B., van der Hoek, W., and Meyer, J.-J. C. (1997). Seeing is believing (and so are hearing and jumping). *Journal of Logic, Language and Information*, pages 33–61.

Wasserman, R. (1999). *Resource-bounded Belief Revision*. PhD thesis, University of Amsterdam, The Netherlands.