

THE UNEXPECTED ASPECTS OF SURPRISE

EMILIANO LORINI

*Institute of Cognitive Sciences and Technologies-CNR, Via San Martino della Battaglia 44,
00185, Roma, ITALY, emiliano.lorini@istc.cnr.it*

CRISTIANO CASTELFRANCHI

*Institute of Cognitive Sciences and Technologies-CNR, Via San Martino della Battaglia 44,
00185, Roma, ITALY, cristiano.castelfranchi@istc.cnr.it*

Abstract

Some symbolic AI models for example BDI (belief, desire, intention) models are conceived as explicit and operational models of the intentional pursuit and belief dynamics. The main concern of these models is to provide a clear understanding of the functional roles of different kinds of epistemic and motivational states (beliefs, acceptances, expectations, intentions, goals, desires etc...), of the relational properties among them. Mental configurations of appraisal (involving different kinds of motivational and epistemic states) which correspond to particular cognitive emotions such as disappointment, fear, relief, shame etc... have been analyzed by several authors close to the BDI theoretical tradition. The main objective of this work is a conceptual and theoretical clarification of the functional role of Surprise in a BDI-like cognitive architecture. In the paper different kinds of surprise are discussed and their properties analyzed. Each type of surprise is associated with a particular phases of the cognitive processing and involves a particular kind of epistemic state (expectation under scrutiny, presupposed belief and so on). A clarification of the functional role of Surprise in a BDI-like cognitive architecture with respect to *resource bounded belief revision* is given.

1 Introduction

Some symbolic AI models and logics for example BDI models and logics ([3, 8, 18, 21]) are conceived as explicit and operational models of the intentional pursuit and belief dynamics. The main concern of these models is to provide a clear understanding of the functional roles of different kinds of epistemic and motivational states (beliefs, acceptances, expectations, intentions, goals, desires etc...), of the relational properties among them. Mental configurations of appraisal (involving different kinds of motivational and epistemic states) which correspond to particular cognitive emotions such as disappointment, fear, relief, shame etc... have been analyzed by several authors close to the BDI theoretical tradition (see for example [5, 16]).

In this work a formal model of *Surprise* based on a BDI-like cognitive architecture and in particular on a theory of expectations ([6, 19]) is developed. In

order to model *Surprise* a propositional logic with probabilities is used. The syntax and semantics of the formal logic is presented in section 2.

In section 3 the process of interpretation of input data will be modeled by defining a abductive procedure for selecting the best interpretation of input data. *Mismatch-based Surprise* will be defined as the the reaction of alert due to a mismatch between an expectation under scrutiny (a belief about the future, that the cognitive system is willing to verify) and the selected interpretation of input data; the *degree of Mismatch-based Surprise* will be associated to the degree of mismatch between the selected interpretation of input data and the invalidated expectation under scrutiny. It will be formally derived that “the higher the probability associated to the invalidated expectation under scrutiny the higher the intensity of the surprise due to its invalidation” and “the lower the probability associated to the explanation invalidating the expectation under scrutiny, the higher the intensity of the surprise due to the invalidation”.

In section 4 we will argue that the revision of beliefs and expectations in the background of the agent’s knowledge structure is associated to a qualitatively different form of surprise. We will focus on the revision of presupposed assumptions and beliefs given the current expectation under scrutiny invalidated by the explanation of input data. *Presupposed beliefs relative to a current expectation under scrutiny* will be defined as beliefs whose contents are implied by the content of the expectation under scrutiny and which are in background during the phase of interpretation of input data. The *degree of Revision-based of surprise* will be associated to the *degree of revision* of a belief presupposed by an invalidated expectation under scrutiny. It will be formally derived that the intensity of *Revision-based of surprise* is always equal or higher than the intensity of *Mismatch-based Surprise*.

In section 5 we will analyze the surprise generated by the *deeper mismatch* between the selected explanation of input data and the (logical) supports of the expectation under scrutiny. A *support of an expectation under scrutiny* will be defined as the background belief whose content implies the content of the expectation under scrutiny. It will be formally derived that the intensity of this deeper *mismatch-based surprises* is always equal or lower than the intensity of *first-order mismatch-based Surprise* (due to the invalidation of the expectation under scrutiny by the selected explanation of input data).

The conclusive part of the work is devoted to discuss the importance of the present approach for the theory of belief revision. We will move in this part toward a procedural perspective on the problem of Surprise. We will argue that *deeper mismatch-based surprises* should be conceived as later emotional responses which are coupled with the revision of deeper beliefs in a belief base. We will defend the following thesis concerning the functional role of surprise in resource bounded cognitive agents: *since realistic cognitive agents are non-omniscient and have not direct and instantaneous access to all their knowledge both in phase of perception and in phase of revision, some mechanism which is responsible :*

1) *for signaling the global inconsistency of the belief base with respect to the incoming input data ,*

- 2) for making explicit deeper layers,
- 3) for the revision of broader parts of the belief base, is needed.

One of the functional role of surprise is exactly this.

Therefore the present work tries to give contributes for the following two scientific areas.

With respect to the area of cognitive modelling the present work provides a conceptual and formal clarification of the notion of Surprise thanks to the elaboration of a typology of surprises where each type of surprise is associated with a particular phase of the cognitive processing and involves particular kinds of epistemic representations (expectation under scrutiny, presupposed belief and so on). Indeed the authors believe that current formal and psychological models of surprise (see for example [2, 17, 15, 20]), which are merely focused on the initial phase of mismatch between input data and active schema and expectation, are incomplete.

With respect to the area of belief revision the present work builds the bases for developing a model of *Surprise-driven progressively structured belief revision* in resource bounded cognitive agents. With respect to this scientific area the present work is very close to works done under the philosophically inspired ([12]) *local revision* framework ([11, 25]) which has been proposed as an advancement of standard *AGM-based belief revision* ([1]) approaches to understand belief revision in resource bounded cognitive agents.

2 Formal foundations

Let us give the formal specification of the model. We define in this paragraph the language that we are adopting with the related syntax and semantic. We use a standard propositional logic with probabilities with a semantic similar to the semantic given in [10].

2.1 Syntax

The primitives of the formal language are the following:

- A set of agent variables $AGT = \{i, j, \dots\}$;
- A set of propositional variables $\Pi = \{p, q, \dots\}$.

The set of *propositional formulas* is defined by the the closure of Π under the Boolean operations \wedge and \neg . Let us use φ and ψ to represent propositional formulas. A *primitive term* is an expression of the form $P_i(\varphi)$ where φ is a propositional formula. A *term* is an expression of the form $a_1 P_i(\varphi_1) + \dots + a_k P_i(\varphi_k)$ where a_1, \dots, a_k are real numbers and $k \geq 1$. A *basic formula* is a statement either of the form $\frac{t}{t'} \geq c$ or of the form $t \geq c$ where t and t' are terms and c is a real number. A *complex formula* is a Boolean combination of *basic formulas*. Let us use f and g to represent complex formulas. Finally a *conditional formula* is a statement of the form $\varphi \Rightarrow_i \psi$ (where φ and ψ are propositional formulas). We use the standard abbreviation for expressing

conditional probabilities $P_i(\varphi|\psi) \geq c =_{\text{def}} \frac{P_i(\varphi \wedge \psi)}{P_i(\psi)} \geq c$. Finally we use the standard notation $\|\varphi\|^S$ to represent the set of worlds in S at which φ is true. Formally $\|\varphi\|^S = \{v \in S \mid M, v \models \varphi\}$.

2.2 Semantics

A model for our logic is defined as $M = (W, X, \pi)$ where:

- W is a non-empty set of possible worlds;
- π is a valuation function which assigns a truth value to propositions in possible worlds $\pi : W \times \Pi \rightarrow B$ where B is the set of boolean variables $B = \{true, false\}$
- X is a function which associates with each world w and agent i AGT a probability space $(W_{w,i}, F_{w,i}, X_{w,i})$ where:
 - $W_{w,i}$ is a subset of W called *sample space*;
 - $F_{w,i}$ is the algebra of subsets built on $W_{w,i}$,
 - $X_{w,i}$ is probability function defined on $2^{W_{w,i}}$ such that $X_{w,i} : 2^{W_{w,i}} \rightarrow [0, 1]$.

2.2.1 Truth conditions and validity

- $M, w \models p$ iff $\pi(p, w) = \text{true}$.
- $M, w \models \top$ for all $w \in W$.
- $M, w \models \neg\varphi$ iff not $M, w \models \varphi$.
- $M, w \models \varphi \wedge \psi$ iff $M, w \models \varphi$ and $M, w \models \psi$.
- $M, w \models \frac{a_1 P_i(\varphi_1) + \dots + a_k P_i(\varphi_k)}{a_{k+1} P_i(\varphi_{k+1}) + \dots + a_l P_i(\varphi_l)} \geq c$ iff $\frac{a_1 X_{w,i}(\|\varphi_1\|^{W_{w,i}}) + \dots + a_k X_{w,i}(\|\varphi_k\|^{W_{w,i}})}{a_{k+1} X_{w,i}(\|\varphi_{k+1}\|^{W_{w,i}}) + \dots + a_l X_{w,i}(\|\varphi_l\|^{W_{w,i}})} \geq c$
where $(\|\varphi_j\|^{W_{w,i}}) = \{v \in W_{w,i} \mid M, v \models \varphi_j\}$
- $M, w \models \varphi \Rightarrow_i \psi$ iff $\forall v \in W_{w,i}$ if $M, v \models \varphi$ then $M, v \models \psi$.

With respect to the axioms of the formal logic we use the axioms given in [10]¹.

¹In [10] the following category of axioms are given. 1) Axioms for propositional reasoning: - Modus Ponens, - Propositional tautologies. 2) Axioms for reasoning about linear inequalities. 3) Axioms for reasoning about probabilities: - $P_i(\varphi) \geq 0$, - $P_i(\text{true}) \geq 0$, - $P_i(\varphi \wedge \psi) + P_i(\varphi \wedge \neg\psi) = P_i(\varphi)$ (additivity), - $P_i(\varphi) = P_i(\psi)$ if $P_i(\varphi) \longleftrightarrow P_i(\psi)$.

2.3 Abducible formulas, input data and expectations under scrutiny

In the present section we define some ad-hoc formal constructions. This implies an extension of our definition of model. Given the following definition 1, 2 and 3 our models are now defined by the tuple $M = (W, X, \pi, \Gamma, \mathbf{Test}, \mathbf{Abd})$.

Definition 1: Perceived data. We define a *perception function* Γ as the function which assigns a set of propositional variables to agents in possible worlds. Formally $\Gamma: AGT \times W \rightarrow 2^{\Pi}$. The input data for an agent i at a given world w is the set $\Gamma_{w,i}$ of propositional variables. $\langle \Gamma \rangle_{w,i}$ is the logical conjunction of propositional variables in $\Gamma_{w,i}$. Formally, $\langle \Gamma \rangle_{w,i} = \bigwedge_{\varphi_i \in \Gamma_{w,i}} \varphi_i$. Thus the elements in $\Gamma_{w,i}$ are the data perceived by the agent i at world w and that the agent must interpret on the basis of his pre-existent belief structure.

Definition 2: Tested expectation (expectation under scrutiny). We also introduce a *test function* \mathbf{Test} which assigns an arbitrary set of propositional variables to agents in possible worlds². Formally $\mathbf{Test}: AGT \times W \rightarrow 2^{\Pi}$. The set of constituents of a tested expectation (expectation under scrutiny) for an agent i at a given world w is the set $\mathbf{Test}_{w,i}$ of propositional variables. $\langle \mathbf{Test} \rangle_{w,i}$ is the logical conjunction of propositional variables in $\mathbf{Test}_{w,i}$ and defines the expectation that the agent i is verifying at a given world w . Formally, $\langle \mathbf{Test} \rangle_{w,i} = \bigwedge_{\varphi_i \in \mathbf{Test}_{w,i}} \varphi_i$. Let us give the semantic for the $\mathbf{Test}_i(\varphi)$ formula to be evaluated at a given world w in W :

$M, w \models \mathbf{Test}_i(\varphi)$ iff $\varphi = \langle \mathbf{Test} \rangle_{w,i}$.

Definition 3: Abducible formulas. We finally introduce an *abducible formula function* \mathbf{Abd} which assigns a power set of propositional variables to agents in possible worlds. Formally $\mathbf{Abd}: AGT \times W \rightarrow 2^{2^{\Pi}}$. $\mathbf{Abd}_{w,i}$ defines the power set of propositional variables given by the function \mathbf{Abd} for agent i and world w . $\mathbf{Abd}_{w,i}^j$ identifies the j -element of $\mathbf{Abd}_{w,i}$ and finally $\langle \mathbf{Abd} \rangle_{w,i}$ is the set of logical conjunctions of propositional variables belonging to the same $\mathbf{Abd}_{w,i}^j$ in $\mathbf{Abd}_{w,i}$. Formally $\langle \mathbf{Abd} \rangle_{w,i} = \left\{ \bigwedge_{\varphi_i \in \mathbf{Abd}_{w,i}^j} \varphi_i \mid \mathbf{Abd}_{w,i}^j \in \mathbf{Abd}_{w,i} \right\}$. Let us give a different notation for elements in $\langle \mathbf{Abd} \rangle_{w,i}$. Let us use H_1, \dots, H_i to represent these elements.

We assume that elements in $\langle \mathbf{Abd} \rangle_{w,i}$ are mutually contradictory.

Assumption 1. If H_i and $H_j \in \langle \mathbf{Abd} \rangle_{w,i}$ then $H_i \wedge H_j \Rightarrow_i \perp$.

²Our *test function* is comparable to the *awareness function* given in [9].

3 Mismatch-based Surprise

3.1 Interpretation of input data

We make the assumption that an agent i at a given world w must interpret input data $\Gamma_{w,i}$ by means of a *classical* abductive procedure of explanation extraction³. Basically the procedure returns the abducible formula with the highest explanatory with respect to the input dataset $\Gamma_{w,i}$. Let us define the procedure.

Definition 4: Procedure of explanation extraction and selection.

1. For each abducible formula $H_i \in \langle Abd \rangle_{w,i}$ calculate the *maximal* subset $max - \Gamma_{w,i}^{H_i}$ of the set of input data $\Gamma_{w,i}$ such that $H_i \Rightarrow_i \langle max - \Gamma \rangle_{w,i}^{H_i}$ where $\langle max - \Gamma \rangle_{w,i}^{H_i} = \bigwedge_{\varphi_i \in max - \Gamma_{w,i}^{H_i}} \varphi_i$.
2. Select the abducible formula $H_i \in \langle Abd \rangle_{w,i}$ with the highest *explanatory value* as the best explanation of the input data.

We introduce a measure of explanatory value where two different parameters are taken into account:

1. the number of perceived data that the abducible formula is able to explain and
2. the prior probability associated to the abducible formula.

With respect to the second parameter we use standard probability theory. The explanatory value of each candidate explanation (abducible formula) H_i depends on to the posterior probability that H_i is true, given the disjunction $H_1 \oplus \dots \oplus H_m$ of candidate explanations (abducible formulas). Using Bayes theorem we obtain:

$$P_i(H_i | H_1 \cup \dots \cup H_m) = \frac{P_i(H_1 \cup \dots \cup H_m | H_i) \cdot P(H_i)}{P_i(H_1 \cup \dots \cup H_m)} \quad (1)$$

Given previous assumption 1 (candidate explanations are mutually contradictory) (1) is equivalent to the following formula.

$$P_i(H_i | H_1 \cup \dots \cup H_m) = \frac{P_i(H_i)}{P_i(H_1) + \dots + P_i(H_m)} \quad (2)$$

³With *classical* abductive procedure (or schema) we mean the most straightforward conception of an inference to an *explanation* of the (logical conjunction of) input data $\langle \Gamma \rangle_{w,i}$: an *explanation* of the input data is some formula H logically entailing $\langle \Gamma \rangle_{w,i}$: $H \Rightarrow_i \langle \Gamma \rangle_{w,i}^{H_i}$. Classical schemas are not the only accepted for abduction. See for example [4, 24] for a detailed analysis of less standard schema for logical abductions.

Moreover the explanatory value of a candidate explanation should depend on the number of data that the hypothesis explains. Therefore the previous probabilistic measure is weighted for the following ratio.

$$\gamma(H_i, \Gamma_{w,i}) = \left(\frac{|max - \Gamma_{w,i}^{H_i}|}{|\Gamma_{w,i}|} \right) \quad (3)$$

where $|\Gamma_{w,i}|$ is the number of elements in $\Gamma_{w,i}$ (cardinality of $\Gamma_{w,i}$), $|max - \Gamma_{w,i}^{H_i}|$ is the maximal number of input data that H_i explains (see definition 4) and κ is an arbitrary constant in $[0, 1]^4$.

Therefore the explanatory value of a candidate explanation H_i with respect to the input dataset $\Gamma_{w,i}$ and given a set of contradictory candidate explanation $\langle Abd \rangle_{w,i} = \{H_1, \dots, H_m\}$ is defined by the following formula which enriches the formal basic language defined in section 2. Semantical conditions of satisfiability are given.

Definition 5a: Explanatory value of an abducible formula.

$$M, w \models \text{Expl-value}_i(H_i) = c \text{ iff } \{H_1, \dots, H_m\} = \langle Abd \rangle_{w,i} \text{ and } \kappa_{w,i} \cdot \gamma(H_i, \Gamma_{w,i}) \cdot \frac{X_{w,i}(|H_i|^{W_{w,i}})}{X_{w,i}(|H_1|^{W_{w,i}}) + \dots + X_{w,i}(|H_m|^{W_{w,i}})} = c$$

Finally we introduce the special formula **selected-expl_i(H_f)** in order to represent the fact that H_f is the abducible formula which has been selected by agent i as the best explanation of the input data. We provide the following semantics.

Definition 5b: Selected explanation.

$$M, w \models \text{selected-expl}_i(H_i) \text{ iff } \{H_1, \dots, H_m\} = \langle Abd \rangle_{w,i} \text{ and } H_i = \arg \max_{H_i \in \langle Abd \rangle_{w,i}} (\kappa \cdot \gamma(H_i, \Gamma_{w,i}) \cdot \frac{X_{w,i}(|H_i|^{W_{w,i}})}{X_{w,i}(|H_1|^{W_{w,i}}) + \dots + X_{w,i}(|H_m|^{W_{w,i}})})$$

Thus a certain abducible formula H_f is selected by agent i at a given world w as the best explanation of the input data if and only H_f is the abducible formula with highest explanatory value with respect to the set of input data.

3.2 Invalidation of the expectation under scrutiny and surprise

The result of the previous selection process is an explanatory item $H_i \in \langle Abd \rangle_{w,i}$ with the highest explanatory value.

If the selected explanation H_i conflicts with the agent's expectation under scrutiny, the agent's expectation is invalidated due to the "logical" mismatch with the incoming interpretation of the input data. This mismatch generates a certain degree of surprise. Consider an agent expecting that φ (and testing this expectation). Since the winning explanation H_i is either a single

⁴We assume that the arbitrary constant κ is a function which assigns for each agent i and world w a *value of reliability of the perceptive source* in the interval $[0, 1]$. Indeed when the agent perceives some input data he can have some doubt about the fact that his sensors are working well. Therefore the explanatory value of the candidate hypothesis is reduced in order to integrate this uncertainty about the reliability of the perceptual apparatus.

propositional atom or a conjunction of propositional atoms, there could be one or more propositional atoms p_1, \dots, p_n of the logical conjunction H_i such that $p_1 \wedge \dots \wedge p_n \wedge \varphi \Rightarrow \perp$. This means that the selected explanation H_i and the expectation that φ are actually conflicting.

We assume that the intensity of superficial Surprise is measured by the degree of mismatch between the expectation under scrutiny and the selected explanation where the degree of mismatch between arbitrary formulas is given by the following formula.

$$\text{Degree-Mismatch}(\varphi, H_i) = \begin{cases} P(\varphi) - P(H_i) & \text{if } P(\varphi) > P(H_i) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

We provide next the basic semantics for the notion of *degree of mismatch* between propositional formulas.

Definition 6: Degree of Mismatch⁵.

$M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) < c$ iff $X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq X_{w,i}(\|H_i\|^{W_{w,i}})$ and $c > 0$ or

$X_{w,i}(\|\varphi\|^{W_{w,i}}) > X_{w,i}(\|H_i\|^{W_{w,i}})$ and $X_{w,i}(\|\varphi\|^{W_{w,i}}) - X_{w,i}(\|H_i\|^{W_{w,i}}) < c$

$M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) = c$ iff $X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq X_{w,i}(\|H_i\|^{W_{w,i}})$ and $c = 0$ or

$X_{w,i}(\|\varphi\|^{W_{w,i}}) > X_{w,i}(\|H_i\|^{W_{w,i}})$ and $X_{w,i}(\|\varphi\|^{W_{w,i}}) - X_{w,i}(\|H_i\|^{W_{w,i}}) = c$

$M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) > c$ iff $X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq X_{w,i}(\|H_i\|^{W_{w,i}})$ and $c < 0$ or

$X_{w,i}(\|\varphi\|^{W_{w,i}}) > X_{w,i}(\|H_i\|^{W_{w,i}})$ and $X_{w,i}(\|\varphi\|^{W_{w,i}}) - X_{w,i}(\|H_i\|^{W_{w,i}}) > c$

Let us give next the basic semantics for Mismatch-based Surprise as the mismatch between an expectation under scrutiny and the invalidating selected explanation of input data.

Definition 7: Mismatch-based Surprise.

$\text{Mismatch-BasedSurprise}_i(\varphi, H_i) = c \stackrel{\text{def}}{=} \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\varphi) \wedge \text{Degree-Mismatch}_i(\varphi, H_i) = c$

$\text{Mismatch-BasedSurprise}_i(\varphi, H_i) < c \stackrel{\text{def}}{=} \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\varphi) \wedge \text{Degree-Mismatch}_i(\varphi, H_i) < c$

$\text{Mismatch-BasedSurprise}_i(\varphi, H_i) > c \stackrel{\text{def}}{=} \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\varphi) \wedge \text{Degree-Mismatch}_i(\varphi, H_i) > c$

Given the previous two definitions we provide the following principles concerning the measure of intensity of surprise obtained by the mismatch between the expectation under scrutiny and the selected explanation of the input data invalidating the expectation.

⁵The second argument of the formula is H_i which is used to represent abducible formulas. However, this notion of degree of mismatch is the most general one and is applicable to every propositional formula of the language.

Proposition 2.1. Given two agents i and j and a value of probability P_1 for agent i 's expectation under scrutiny, a value of probability P_2 for agent j 's expectation under scrutiny, a value of probability P_3 for agent i 's explanation of input data and a value of probability P_4 for agent j 's explanation of input data, if $P_1 > P_2$ and $P_3 = P_4$ then agent i 's feels more (or equally) surprised than agent j due the invalidation of the expectation under scrutiny. Formally:

$$\models (P_i(\varphi) > P_j(\psi)) \wedge (P_i(H_f) = P_j(H_g)) \wedge \text{Mismatch-BasedSurprise}_j(\psi, H_g) = c \wedge \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\varphi) \rightarrow (\text{Mismatch-BasedSurprise}_i(\varphi, H_f) > c)$$

Proposition 2.2. Given two agents i and j and a value of probability P_1 for agent i 's expectation under scrutiny, a value of probability P_2 for agent j 's expectation under scrutiny, a value of probability P_3 for agent i 's explanation of the input data and a value of probability P_4 for agent j 's explanation of input data, if $P_1 = P_2$ and $P_3 < P_4$ then agent i 's feels more surprised than agent j due the invalidation of the expectation under scrutiny. Formally:

$$\models (P_i(\varphi) = P_j(\psi)) \wedge (P_i(H_f) < P_j(H_g)) \wedge \text{Mismatch-BasedSurprise}_j(\psi, H_g) = c \wedge \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\varphi) \rightarrow (\text{Mismatch-BasedSurprise}_i(\varphi, H_f) > c)$$

Proof. The proofs of Prop. 2.1 and 2.2 are obtained straightforward from the definition 6 of degree of mismatch and definition 7 ■

The intuitive reading of the previous propositions 2.1 and 2.2 is that “the higher the probability associated to the invalidated expectation under scrutiny the higher the intensity of the surprise due to its invalidation” and “the lower the probability associated to the explanation invalidating the expectation under scrutiny, the higher the intensity of the surprise due to the invalidation”. Proposition 2.1 and 2.2 are reasonable. Indeed intensity of surprise should depend at least on the previous two parameters: the probability of my invalidated expectation and the probability of the interpreted input data. We are very surprised when we are almost certain that a certain fact will happen and our expectation gets invalidated. We are even more surprised when our expectation is invalidated by something that we perceive and that we believe to be almost impossible.

Another principle looks also reasonable. Since surprise is the response to a mismatch between two representations (the anticipated and expected one and the explanation input data), and since a mismatch not necessarily is a yes/no result but can be partial (given that the representation is a pattern of features), one might claim that: *The higher is the number of mismatching features, the stronger the surprise.* This principle is too simplistic and naive . Let us argue

against it by the following example.

Example. Consider an agent driving with the car in a street. The agent is ready to stop at the next traffic light since he *expects* the next traffic light to be *red and not green*. The agent is actively testing his expectation that $(r \wedge \neg v)$. Moreover the agent assigns the following conditional probabilities

| $P(V R)$ | v | $\neg v$ |
|----------|------|----------|
| r | 0,05 | 0,95 |
| $\neg r$ | 1 | 0 |

and the following probabilities to the event “the traffic light will be red” and the event “the traffic light will not be red”

| | r | $\neg r$ |
|--------|-----|----------|
| $P(R)$ | 0,8 | 0,2 |

Consider now two different situations.

1) Imagine that the agent perceives the traffic light to be *not red and green* ($v \wedge \neg r$). The input data invalidate the expectation that $(r \wedge \neg v)$ and more than this they mismatch with two out of two propositional features of the expected representation: both propositional atoms in the (expected) conjunction $r \wedge \neg v$ are invalidated by the input data.

2) Imagine that the agent perceives the traffic light to be *red and green* ($v \wedge r$). Again the input data invalidate the expectation that $(r \wedge \neg v)$ but only mismatch with one out of two propositional features of the expected representation: only the propositional atom $\neg v$ in the (expected) conjunction $r \wedge \neg v$ is invalidated by the input data.

Let us calculate the degree of mismatch (and consequently the intensity surprise) in situations 1 and 2 and compare them.

$$(i) \text{ Degree-Mismatch } (r \wedge \neg v, v \wedge \neg r) = (P(\neg v|r) \cdot P(r)) - (P(v|\neg r) \cdot P(\neg r)) = 0.76 - 0.2 = 0.56$$

$$(ii) \text{ Degree-Mismatch } (r \wedge \neg v, v \wedge r) = (P(\neg v|r) \cdot P(r)) - (P(v|r) \cdot P(r)) = 0.76 - 0.04 = 0.72.$$

This example shows that not necessarily *higher the number of mismatching features, more intense is the surprise*.

In fact, the features composing the expected and tested representation as well as the features composing the representation explaining the input data are generally dependent one from each other. Thus if the features have some degree of interdependence, if there is some internal coherence of the global representation, one might be more surprised when one feature F1 of the global representation explaining the input data (which invalidates the tested expectation) is evaluated to be very unlikely given feature F2 (in order to express the degree of interdependence among the features we use conditional probabilities).

4 First layer revision and crisis of presupposed assumptions

After the invalidation of the expectation under scrutiny by the explanation of input data the agent needs to revise the first layer of his knowledge structure. The revision of beliefs and expectations in the background of the agent’s belief structure is associated to a qualitatively different form of surprise. Let us focus in this section on the revision of presupposed assumptions and beliefs relative to the current expectation under scrutiny invalidated by the explanation of input data. A *presupposed belief relative to an expectation under scrutiny* is a belief whose content is implied by the content of the expectation under scrutiny and which were in background during the phase of interpretation of the input data. Therefore in our model a *presupposed belief* is always given with respect to some *expectation or belief under scrutiny*. Our definition of presupposed belief is slightly different from the definition of presupposition given in speech act theory ([22]). According to speech act theory when agent i says “the strawberry is red! Is it?” agent i is asserting or questioning about the colour of the fruit and is just presupposing that “the strawberry is coloured”. The same presupposition holds when agent i says “the strawberry is not red!”. Therefore in speech act theory a given belief that A presupposes a belief that B if and only if the agent must believe that B in order either to believe that A is true or to believe that A is false.

Definition 8: Belief that ψ presupposed by tested expectation that φ .
 Presupposed $_i$ (ψ, φ) =_{def} Test $_i$ (φ) \wedge ($\varphi \Rightarrow_i \psi$)

It is well known that the invalidation of the *asserted and tested* belief can be surprising; but the invalidation of the presupposed belief can be surprising as well. We argue that:

the invalidation of tested expectation generates a mismatch-based surprise with a certain intensity and requires a restructuring of the unquestioned assumptions which in turn generates a revision-based of surprise.

Let us first all give a quantitative definition of *Revision-based of surprise*. We will move afterwards to analyze the surprise associated with the revision of presupposed beliefs of a given tested expectation which are invalidated by the explanation of input data. We will finally provide a comparison of the intensity of this form of surprise with the intensity of the mismatch-based surprise associated with the invalidation of the tested expectation implying the presupposed assumption (see previous section 3).

Let us introduce a general definition of degree of revision which is given with respect to arbitrary formulas.

$$\text{Degree-Revision} (\varphi, H_i) = \begin{cases} P(\varphi) - P(\varphi|H_i) & \text{if } P(\varphi) > P(\varphi|H_i) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

We provide next the basic semantics of special degree-Revision formulas.

Definition 9: Degree of Revision.

$$\begin{aligned}
M, w \models \text{Degree-Revision}_i(\varphi, H_i) < c &\text{ iff } X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} \text{ and } \\
c > 0 &\text{ or} \\
X_{w,i}(\|\varphi\|^{W_{w,i}}) > \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} &\text{ and } X_{w,i}(\|\varphi\|^{W_{w,i}}) - \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} < c \\
M, w \models \text{Degree-Revision}_i(\varphi, H_i) = c &\text{ iff } X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} \text{ and } c \\
= 0 &\text{ or} \\
X_{w,i}(\|\varphi\|^{W_{w,i}}) > \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} &\text{ and } X_{w,i}(\|\varphi\|^{W_{w,i}}) - \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} = c \\
M, w \models \text{Degree-Revision}_i(\varphi, H_i) > c &\text{ iff } X_{w,i}(\|\varphi\|^{W_{w,i}}) \leq \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} \text{ and } \\
c < 0 &\text{ or} \\
X_{w,i}(\|\varphi\|^{W_{w,i}}) > \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} &\text{ and } X_{w,i}(\|\varphi\|^{W_{w,i}}) - \frac{X_{w,i}(\|H_i \wedge \varphi\|^{W_{w,i}})}{X_{w,i}(\|H_i\|^{W_{w,i}})} > c
\end{aligned}$$

Let us give next the basic semantics for Revision-based Surprise (ψ, H_i, φ) as the degree of revision $\text{Degree-Revision}_i(\psi, H_i)$ where φ is the content of the expectation under scrutiny, ψ is the content of some belief presupposed by the expectation that φ , H_i is the selected explanation of input data and H_i invalidates both φ and ψ .

Definition 10: Revision-based Surprise.

$$\begin{aligned}
\text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c &=_{\text{def}} \text{Presupposed}_i(\psi, \varphi) \wedge \text{selected-} \\
\text{expl}_i(H_i) \wedge (H_i \Rightarrow_i \neg\psi) \wedge \text{Degree-Revision}_i(\psi, H_i) = c \\
\text{Revision-basedSurprise}_i(\psi, H_i, \varphi) > c &=_{\text{def}} \text{Presupposed}_i(\psi, \varphi) \wedge \text{selected-} \\
\text{expl}_i(H_i) \wedge (H_i \Rightarrow_i \neg\psi) \wedge \text{Degree-Revision}_i(\psi, H_i) > c \\
\text{Revision-basedSurprise}_i(\psi, H_i, \varphi) < c &=_{\text{def}} \text{Presupposed}_i(\psi, \varphi) \wedge \text{selected-} \\
\text{expl}_i(H_i) \wedge (H_i \Rightarrow_i \neg\psi) \wedge \text{Degree-Revision}_i(\psi, H_i) < c
\end{aligned}$$

On the basis of the previous formulation the intensity of Revision-based surprise is:

- 1) the numeric distance between the prior probability associated to the presupposed belief that ψ and the conditional probabilities that ψ is true given the selected explanation of input data H_i (invalidating ψ) when this distance is positive;
- 2) null when the distance is null or a negative number.

Basically, Revision-based surprise measures the magnitude of the restructuring decrease of probability of a certain presupposed belief that ψ given the selected explanation of input data H_i invalidating ψ .

From the previous definition of Revision-based surprise we derive the following principle.

Proposition 3. The surprise associated with the revision of a presupposed belief that ψ invalidated by the explanation of input data H_i is equal to the prior probability of ψ . Formally:

$$\models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \longleftrightarrow P_i(\psi) = c$$

Proof. Let us prove only one direction of the equivalence. We only prove that $\models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) = c$. If $M, w \models H_i \wedge \psi \Rightarrow_i \perp$ then $M, w \models P_i(\psi \wedge H_i) = 0$. Therefore $M, w \models P_i(\psi|H_i) = \frac{P_i(\psi \wedge H_i)}{P_i(H_i)} = 0$. From the previous equivalence it follows that: $M, w \models P_i(\psi) \geq P_i(\psi|H_i) = 0$. From definition 10 we have: 1) $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Revision}_i(\psi, H_i) = c$. From the semantics of degree of revision given in definition 9 we can establish that 2) $M, w \models \text{Degree-Revision}_i(\psi, H_i) = c \rightarrow ((P_i(\psi) > P(\psi|H_i) \rightarrow c = P_i(\psi) - P(\psi|H_i)) \wedge (P_i(\psi) \leq P(\psi|H_i) \rightarrow c = 0))$. Since $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow (H_i \wedge \psi \Rightarrow_i \perp)$ we can conclude that either 3a) $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) > P_i(\psi|H_i) = 0$ or 3b) $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) = P_i(\psi|H_i) = 0$. Let us distinguish two cases.

CASE 1. Assume that $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) > P_i(\psi|H_i) = 0$. Therefore from 1) and 2) we have that $\text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow c = P_i(\psi)$.

CASE 2. Assume that $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) = P_i(\psi|H_i) = 0$. Therefore from 1) and 2) we have that $\text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow c = P_i(\psi) = 0$.

This is enough to conclude that: $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\psi) = c$. The opposite direction of the equivalence can similarly be proved. ■

Let us show now that the invalidation of a presupposed belief implies the invalidation of the expectation under scrutiny which presupposes it. Therefore the surprise associated to the revision of an invalidated and presupposed belief is always added to the surprise associated to the mismatch between the explanation of input data and the presupposing expectation under scrutiny (invalidated by the explanation). The previous intuition is confirmed by the following proposition.

Proposition 4. Given H_i the explanation of input data, φ the content of an expectation under scrutiny and ψ the content of a belief presupposed by the previous expectation that φ , if the presupposed belief that ψ is invalidated by H_i then the expectation that φ is also invalidated by H_i . Formally:

$$\models (\text{Presupposed}_i(\psi, \varphi) \wedge H_i \Rightarrow_i \neg\psi) \rightarrow (H_i \Rightarrow_i \neg\varphi)$$

Proof. From $M, w \models H_i \Rightarrow_i \neg\psi$ (the presupposed belief that ψ is invalidated by H_i) and $M, w \models \varphi \Rightarrow_i \psi$ (ψ is presupposed by φ) which is equivalent to $M, w \models \neg\psi \Rightarrow_i \neg\varphi$ it follows that $M, w \models H_i \Rightarrow_i \neg\varphi$ ■

In the next proposition 5 we compare Mismatch-based Surprise with Revision-based Surprise.

Proposition 5. Given an expectation that φ under scrutiny and its presupposed belief that ψ , if the latter is invalidated by the explanation of input data

H_i then the surprise associated with the invalidation of the expectation that φ under scrutiny is equally or less intense than the surprise associated with the revision of the presupposed belief that ψ . Formally:

$$\models \text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Mismatch-basedSurprise}_i(\varphi, H_i) \leq c$$

Proof. From proposition 4, definition 7 and definition 10 it follows that $M, w \models \text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow (H_i \Rightarrow_i \neg\varphi)$. From definition 7 and 10 it follows that $M, w \models \text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_i)$. We must now prove that $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Mismatch}_i(\varphi, H_i) \leq c$. Let us assume that $\text{Degree-Mismatch}_i(\varphi, H_i) = d$. We must prove that $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Mismatch}_i(\varphi, H_i) = d \wedge (d \leq c)$.

From definition 6 we have that 1) either $\text{Degree-Mismatch}_i(\varphi, H_i) = d$ and $d = P_i(\varphi) - P_i(H_i)$ or $\text{Degree-Mismatch}_i(\varphi, H_i) = d$ and $d = 0$. Let us consider proposition 3 and reformulate it as 2) $M, w \models \text{Revision-basedSurprise}_i(\psi, H_i, \varphi) = P_i(\psi)$. Finally let us make explicit the following fact. It holds (from definitions 8 and 10) that $M, w \models \text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow (\varphi \Rightarrow_i \psi)$. Moreover it holds (from the semantic of $\varphi \Rightarrow_i \psi$) that $M, w \models (\varphi \Rightarrow_i \psi) \rightarrow P_i(\varphi) \leq P_i(\psi)$. Therefore 3) $M, w \models \text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow P_i(\varphi) \leq P_i(\psi)$.

Therefore from 1), 2) and 3) (and the obvious fact $P_i(\psi) \geq 0$) we can conclude that $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Mismatch}_i(\varphi, H_i) = d \wedge (d \leq c)$ and finally we can conclude that $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Mismatch}_i(\varphi, H_i) \leq c$. From: - $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Test}_i(\varphi) \wedge \text{selected-expl}_i(H_i)$, - $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Degree-Mismatch}_i(\varphi, H_i) \leq c$ and - the definition 7 of Mismatch-based surprise we conclude that $\text{Revision-BasedSurprise}_i(\psi, H_i, \varphi) = c \rightarrow \text{Mismatch-basedSurprise}_i(\varphi, H_i) \leq c$ ■

For concluding this section let us give an example in order to show the difference among the surprise associated to the invalidation of the expectation under scrutiny and surprise due to the revision of the invalidated belief presupposed by the expectation under scrutiny.

Example. Imagine that agent i expects that “John is married”. Given this expectation agent i has the presupposed belief that “John is an adult”. Agent i wants to test the validity of his expectation and asks to agent j whether “John is married or not”. Agent j gives the following answer “John is not married, he is a child!”. The first part of the answer (“John is not married”) invalidates agent i ’s expectation that “John is married”. Thus agent i feels surprised given the mismatch between his expectation and the incoming input. The second part of the answer (“John is a child”) invalidates the presupposed belief that “John is an adult”. Therefore given the revision of the presupposed belief that “John is

an adult”, agent i feels a secondary surprise. This latter surprise is more intense than the former.

5 Deeper surprises and crisis of doxastic supports

In this section we analyze the surprise generated by the *deeper mismatch* between the selected explanation of input data and the (logical) supports of the expectation under scrutiny. A *support of an expectation under scrutiny* is defined as the background belief whose content implies the content of the expectation under scrutiny.

Example. Imagine that agent i has explained the input data by the fact “there is a car close to me”. The explanation “there is a car close to me” has invalidated the expectation under scrutiny “there is a table close to me” (and perhaps some presupposed belief such as “there is a wooden object close to me”) and generated a mismatch-based surprise with a certain intensity. The agent tries now to answer to the question: - “why am I expecting to perceive a table?” by looking for the supports of the fact “there is a table close to me”, i.e. by looking for some fact which implies the fact “there is a table close to me”. Imagine that the agent explains the fact “there is a table close to me” by the fact “I am in a dining room”. The fact “there is a car close to me” invalidates the first-order support “I am in a dining room” of the expectation “there is a table close to me” and a second-order mismatch-based surprise is generated by the invalidation.

Let us define the notion of support of an expectation under scrutiny as the background belief whose content implies the content of the expectation under scrutiny.

Definition 11: Support ψ of an expectation under scrutiny that φ .
 $\text{Support}_i(\psi, \varphi) =_{\text{def}} \text{Test}_i(\varphi) \wedge (\psi \Rightarrow_i \varphi)$

It is easy to prove the following proposition.

Proposition 6. If the selected explanation of input data invalidates the expectation under scrutiny then it invalidates all supports of this expectation. Formally:

$$\models (\text{Support}_i(\psi, \varphi) \wedge (H_{i \Rightarrow_i} \neg \varphi)) \rightarrow (H_{i \Rightarrow_i} \neg \psi)$$

Proof. It follows straightforward from definition 11. Indeed $\psi \Rightarrow_i \varphi$ in formula $\text{Support}_i(\psi, \varphi)$ can be rewritten as $\neg \varphi \Rightarrow_i \neg \psi$. Therefore, if $M, w \models H_{i \Rightarrow_i} \neg \varphi$ then $M, w \models H_{i \Rightarrow_i} \neg \psi$.

Let us now introduce the notion of Deep Mismatch-based Surprise: the surprise associated to the invalidation of the support of an expectation under scrutiny.

Definition 12: Deep Mismatch-based Surprise.

$\text{DeepMismatch-BasedSurprise}_i(\psi, H_i, \varphi) = c \stackrel{\text{def}}{=} \text{Support}_i(\psi, \varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\psi) \wedge \text{Degree-Mismatch}_i(\psi, H_i) = c$

$\text{DeepMismatch-BasedSurprise}_i(\psi, H_i, \varphi) > c \stackrel{\text{def}}{=} \text{Support}_i(\psi, \varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\psi) \wedge \text{Degree-Mismatch}_i(\psi, H_i) > c$

$\text{DeepMismatch-BasedSurprise}_i(\psi, H_i, \varphi) < c \stackrel{\text{def}}{=} \text{Support}_i(\psi, \varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_f \Rightarrow_i \neg\psi) \wedge \text{Degree-Mismatch}_i(\psi, H_i) < c$

We can now prove the following property relating *mismatch-based surprise* (generated by mismatch between the expectation under scrutiny and the invalidating selected explanation of the input data) with *deep mismatch-based surprise* (generated by the mismatch between a support of the expectation under scrutiny and the invalidating selected explanation of input data).

Proposition 7. Given an expectation that φ under scrutiny and the belief that ψ which supports φ , if the former is invalidated by the selected explanation of input data H_i then the surprise associated with the mismatch between the belief ψ and the invalidating selected explanation H_i is equally or less intense than the surprise associated with the mismatch between the expectation that φ under scrutiny and the invalidating selected explanation H_i . Formally:

$$\models \text{Mismatch-BasedSurprise}_i(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{DeepMismatch-basedSurprise}_i(\psi, H_i, \varphi) \leq c$$

Proof. We need first of all to prove that $M, w \models \text{Mismatch-BasedSurprise}_i(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{Test}_i(\varphi) \wedge (\psi \Rightarrow_i \varphi) \wedge \text{selected-expl}_i(H_f) \wedge (H_i \Rightarrow_i \neg\psi)$ where $\text{Test}_i(\varphi) \wedge (\psi \Rightarrow_i \varphi) \wedge \text{selected-expl}_i(H_i) \wedge (H_f \Rightarrow_i \neg\psi)$ is part of the definition of $\text{DeepMismatch-basedSurprise}_i(\varphi, H_i) \geq c$ (see definition 11 and definition 12). This is indeed the case thank to definition 7 of Mismatch-Based Surprise and proposition 6 which guarantees that $M, w \models (\text{Support}_i(\psi, \varphi) \wedge (H_i \Rightarrow_i \neg\varphi)) \rightarrow (H_i \Rightarrow_i \neg\psi)$.

Afterwards we need to prove that $M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) = c \wedge (\psi \Rightarrow_i \varphi) \rightarrow \text{Degree-Mismatch}_i(\psi, H_i) \leq c$. Let us prove it. From the semantic of degree of mismatch (definition 6) we can state that 1) $M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) = c \wedge (\psi \Rightarrow_i \varphi) \rightarrow ((P_i(\varphi) > P_i(H_i) \rightarrow c = P_i(\varphi) - P_i(H_i)) \wedge (P_i(\varphi) \leq P_i(H_i) \rightarrow c = 0))$. Moreover, from $M, w \models (\psi \Rightarrow_i \varphi)$ it follows that 2) $M, w \models P_i(\psi) \leq P_i(\varphi)$. Let us assume that $M, w \models \text{Degree-Mismatch}_i(\psi, H_i) = d$. Therefore in order to prove that $M, w \models \text{Degree-Mismatch}_i(\varphi, H_i) = c \wedge (\psi \Rightarrow_i \varphi) \rightarrow \text{Degree-Mismatch}_i(\psi, H_i) \leq c$ we must simply prove that $M, w \models d \leq c$. Again from definition 6 we can state that 3) $M, w \models \text{Degree-Mismatch}_i(\psi, H_i) = d \rightarrow ((P_i(\psi) > P_i(H_i) \rightarrow d = P_i(\psi) - P_i(H_i)) \wedge (P_i(\psi) \leq P_i(H_i) \rightarrow d = 0))$. From 1), 2) and 3) it follows $M, w \models$ that $d \leq c$. Let us show why this is the case.

CASE 1. Assume that $M, w \models P_i(\varphi) \leq P_i(H_i)$. Therefore $M, w \models c = 0$ and since $M, w \models P_i(\psi) \leq P_i(\varphi)$ we can conclude that also $M, w \models P_i(\psi) \leq P_i(H_i)$.

Thus $M, w \models d = 0$.

CASE 2. Assume that $M, w \models P_i(\varphi) > P_i(H_i)$. Therefore $M, w \models c = P_i(\varphi) - P_i(H_i)$ and since $M, w \models P_i(\psi) \leq P_i(\varphi)$ we can only conclude that either $M, w \models P_i(\psi) > P_i(H_i)$ or $M, w \models P_i(\psi) \leq P_i(H_i)$. Let us split CASE 2 into two parts.

CASE 2a. If $M, w \models P_i(\psi) > P_i(H_i)$ then $M, w \models d = P_i(\psi) - P_i(H_i)$ and since $M, w \models P_i(\psi) \leq P_i(\varphi)$ we can conclude that $M, w \models d \leq c$.

CASE 2b. If $M, w \models P_i(\psi) \leq P_i(H_i)$ then $M, w \models d = 0$ and since $M, w \models P_i(\varphi) > P_i(H_i)$ and $M, w \models c = P_i(\varphi) - P_i(H_i)$ we can conclude that $M, w \models d \leq c$.

Having proved that $M, w \models d \leq c$, we can conclude that $M, w \models \text{Degree-Mismatch}_1(\varphi, H_i) = c \wedge (\psi \Rightarrow_i \varphi) \rightarrow \text{Degree-Mismatch}_1(\psi, H_i) \leq c$. Finally since $M, w \models \text{Mismatch-BasedSurprise}_1(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{Degree-Mismatch}_1(\varphi, H_i) = c \wedge (\psi \Rightarrow_i \varphi)$ we can conclude that $M, w \models \text{Mismatch-BasedSurprise}_1(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{Degree-Mismatch}_1(\psi, H_i) \leq c$. From $M, w \models \text{Mismatch-BasedSurprise}_1(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{Degree-Mismatch}_1(\psi, H_i) \leq c$ and $M, w \models \text{Mismatch-BasedSurprise}_1(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{Test}_1(\varphi) \wedge (\psi \Rightarrow_i \varphi) \wedge \text{selected-expl}_1(H_f) \wedge (H_f \Rightarrow_i \neg\psi)$ and definition 11 we can conclude that $M, w \models \text{Mismatch-BasedSurprise}_1(\varphi, H_i) = c \wedge \text{Support}_i(\psi, \varphi) \rightarrow \text{DeepMismatch-basedSurprise}_1(\psi, H_i, \varphi) \leq c$ ■

6 Conclusion

The conclusive part of the work is devoted to discuss the importance of the present approach for the theory of belief revision. We are not going to give any formal emphasis on this part. The issues raised in this conclusion are indeed the subjects of other our works ([7]).

Let us move toward a procedural perspective and conceive *deeper mismatch-based surprises* as deeper emotional responses which are coupled with the revision of deeper beliefs in a belief base. We want to defend here the following thesis concerning the functional role of surprise in resource bounded cognitive agents: *since realistic cognitive agents are non-omniscient and have not direct and instantaneous access to all their knowledge both in phase of perception and in phase of revision, some mechanism which is responsible :*

- 1) *for signaling the global inconsistency of the belief base with respect to the incoming input data ,*
- 2) *for making explicit deeper layers,*
- 3) *for the revision of broader parts of the belief base, is needed.*

One of the functional role of surprise is exactly this.

We assume the existence of certain threshold Δ of *acceptable surprise* which has the function of determining whether a given agent decides to continue to verify the consistency of his beliefs and expectations and to revise deeper layers of his

belief base or whether the agent decides to stop with his check of consistency and progressive revision.

If an agent has felt a very intense *first-order mismatch-based surprise* (i.e. the *degree of first-order mismatch-based surprise* is above threshold Δ) due to the invalidation of the expectation under scrutiny by the selected explanation of input data then: 1) the agent locally revises the elements of his belief base which are entailed by the expectation under scrutiny, 2) the agent passes to check the adequacy of deeper layers of his belief base with respect to the incoming input data. This check consists in making explicit the first-order supports of the expectation under scrutiny and is responsible for the generation of a *second-order mismatch-based surprise*. A *second-order mismatch-based surprise* is coupled now with a local belief revision: the agent revises the elements of his belief base which are entailed by the first-order supports of the expectation under scrutiny. If again the *second-order mismatch-based surprise* is very intense (i.e. its degree is above threshold Δ) the agent passes to check the adequacy of deeper layers. He makes explicit the direct supports of the first-order supports of the expectation under scrutiny (which are second-order supports of the expectation under scrutiny). When deeper supports are made explicit a *third-order mismatch-based surprise* is generated. This surprise is coupled with a new local belief revision: the revision of the elements of his belief base which are entailed by logical the second-order supports of the expectation under scrutiny. Deeper and deeper beliefs (deeper and deeper supports of the expectation under scrutiny) are made explicit and local revision extended until the moment at which the intensity of the *n-order mismatch-based surprise* is not intense anymore (i.e. its degree is below threshold Δ). Previous proposition 7 tells us that *mismatch-based surprises* decreases along the top down path going toward deeper layers of the belief base (and toward deeper logical supports of the expectation under scrutiny). Given proposition 7 we can predict that *degree of n-order mismatch-based surprise* $>$ *degree of (n+1)-order mismatch-based surprise* and therefore we can predict that there will be an instant at which the agent will “come out” of the progressive revision of its belief base (the existence of a *stop condition* is guaranteed).

Finally let us consider another relevant and possible extension of our model. Some existing psychological model of Surprise (see [17] for example) establishes that intense Surprise triggers a search for causal explanations, justifications and confirmations of the unexpected event. After that an intense surprise has been felt, the agent tries to cope with it by trying to justify the unexpected event, by trying to confirm it (see also Lazarus’ idea of *Secondary Appraisal* [13]). In the present work the problem of *coping strategies* is left unspecified. We have assumed that the interpretation of input data is one-step process and that once the best explanation of input data has been the selected a gradual revision of the belief structure starts. In order to deal with coping strategies we would need to develop our theory into the following direction. Similarly to Shanahan’s approach ([23]) we should integrate in our model a *feedback mechanism* for readjusting the explanatory values of the candidate explanations (abducible

formulas) . We should model this feedback mechanism in terms of epistemic actions (see also [14]). After that the agent has interpreted the input data, if the selected explanation is very surprising, the agent either: 1) can try to enlarge the set of abducible formulas in order to find a less surprising explanation of input data (an explanation whose explanatory value is higher than the explanatory value of the previously selected and very surprising explanation) or; 2) given the set of abducible formulas he can try to check (by doing some epistemic action for verification) whether the interpretation selected at the first round is the right one.

Example. Imagine that a doctor is almost sure that the patient has a certain pathology x . The doctor decides to subject the patient to a rigorous test in order to confirm the diagnosis. In our model the abducible formulas would be two: “the patient has pathology x ”, “the patient has not pathology x ”. Imagine the result of the test is negative. The doctor interprets this result as “the patient has not pathology x ”. The doctor gets surprised since the expectation that “the patient has pathology x ” under scrutiny is invalidated. How could the doctor cope with this surprise? The doctor could make some epistemic action for verification (before accepting the fact that “the patient has not pathology x ” and revising his beliefs and expectations). For instance the doctor could make some further medical tests.

Acknowledgment

This research has been supported by the European Project MindRACES. Moreover, we would like to thank our colleagues Rino Falcone, Luca Tummolini, Maria Miceli and also Mark Bickhard, Andrew Ortony and the other participants in the panel discussion about Surprise at the AAAI Fall Symposium “From Reactive to Anticipatory Cognitive Embodied Systems”.

References

- [1] Alchourrón, C., Gärdenfors, P., Makinson, D. (1985). On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *The Journal of Symbolic Logic*, 50, pp. 510-530.
- [2] Baldi, P. (2004). Surprise: A Shortcut for Attention? In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of Attention*, Academic Press.
- [3] Bratman, M. E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.
- [4] Boutilier, C., Becher, V. (1995). Abduction as Belief Revision. *Artificial Intelligence*, 77, pp. 43-94.

- [5] Castelfranchi, C. (2000). Affective Appraisal versus Cognitive Evaluation in Social Emotions and Interactions. In A. Paiva (Eds.), *Affective Interactions: Towards a New Generation of Computer Interfaces*, Springer-Verlag, Berlin, pp. 76-106.
- [6] Castelfranchi, C., Lorini, E. (2003). Cognitive Anatomy and Functions of Expectations. In *Proceedings of IJCAI 03 Workshop on Cognitive modeling of agents and multi-agent interaction*, Acapulco, Mexico.
- [7] Castelfranchi, C., Lorini, E. (under review). Surprise and Local Belief Revision. Submitted for *Topoi*.
- [8] Cohen, P. R. , Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, 42, pp. 213-261.
- [9] Fagin, R., Halpern, J. Y. (1987). Belief, Awareness, and Limited Reasoning. *Artificial Intelligence* 34(1), pp. 39-76.
- [10] Halpern, J. Y. (2003). *Reasoning about uncertainty*. MIT Press, Cambridge, MA.
- [11] Hansson, S. O., Wassermann, R. (2002). Local Change. *Studia Logica*, 70(1), pp. 49-76.
- [12] Harman, G. (1986). *Change in View*. MIT Press, Cambridge, MA.
- [13] Lazarus, R.S. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- [14] Lorini, E., Castelfranchi, C. (2004). The role of epistemic actions in expectations. In *Proceedings of the Second Workshop of Anticipatory Behavior in Adaptive Learning Systems (ABIALS 2004)*, Los Angeles, USA, pp. 62-72.
- [15] Macedo, L., Cardoso, A. (2001). Modelling Forms of Surprise in an Artificial Agent. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp. 588-593
- [16] Meyer, J. J. (2004). Reasoning about emotional agents. In *Proceedings of 16th European Conference on Artificial Intelligence*, IOS Press, pp. 129-133.
- [17] Meyer, W. U., Reisenzein, R., Schützwohl, A. (1997). Towards a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21, pp. 251-274.
- [18] Meyer, J.J. Ch., van der Hoek, W., van Linder, B. (1999). A Logical Approach to the Dynamics of Commitments. *Artificial Intelligence*, 113(1-2), pp. 1-40.
- [19] Miceli, M., Castelfranchi. The Mind and the Future. The (Negative) Power of Expectations. *Theory & Psychology*, 12(3), pp. 335-366, 2002.
- [20] Ortony, A., Partridge (1987). Surprisingness and expectation failure: Whats the difference? In *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, pp. 106-108, Los Altos, CA: Morgan Kaufmann.
- [21] Rao, A. S., Georgeff M. P. (1991). Modelling rational agents within a BDI-architecture. In *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning*, Morgan Kaufmann Publishers, San Mateo, CA.

- [22] Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.
- [23] Shanahan, M. (2002). A logical account of perception incorporating feedback and expectation. In *Proceedings of 8th international conference on principles of knowledge representation and reasoning*, Morgan Kaufmann Publishers, pp. 3-13.
- [24] Walliser, B., Zwirn, D., Zwirn, H. (2005). Abductive logics in a belief revision framework. *Journal of Logic, Language and Information*, 14, pp. 87-117.
- [25] Wasserman, R. (1999). *Resource-bounded Belief Revision*. PhD thesis, University of Amsterdam, The Netherlands.