Logique & Analyse **230** (2015), 251-277

THE DYNAMICS OF SURPRISE

LORENZ DEMEY*

1. Introduction

The phenomenon of surprise is ubiquitous in everyday life. People get surprised all the time; for example, by an unexpected flash of light, or—more 'down to earth'—about the fact that their local grocery store has run out of milk (after all, the store is usually well-stocked!). The role of surprise in human life has been intensively studied in psychology from cognitive, social, developmental and educational perspectives. Furthermore, computer scientists have implemented the psychological findings about human surprise in artificial agents, and used logical models to describe these agent architectures. Surprise even crops up in various philosophical debates, such as those concerning the role of surprising evidence in Bayesian epistemology, or concerning the so-called surprise examination paradox.[1]

The overarching goal of this paper is to provide a new analysis of the phenomenon of surprise in the framework of probabilistic dynamic epistemic logic. This account is based on the vast amount of experimental work on surprise in psychology, which should benefit its empirical adequacy. The paper's main thesis, however, is of a more conceptual nature: surprise is an essentially *dynamic* phenomenon, and any good formal analysis should represent this dynamics explicitly. I will argue that all current formalizations of surprise in artificial intelligence and logic fail to fully capture this dynamics, and show that the framework developed in this paper *is* able to capture it. As an additional benefit, this new framework can be used to analyze some aspects of surprise that could not be analyzed before.

* Earlier versions of this paper were presented at the LIRa seminar (ILLC, Amsterdam, October 2012), the Reasoning Club PhD conference (Brussels, September 2012), LOFT 10 (Sevilla, June 2012) and a workshop on modal logic (Brussels, May 2012). I would like to thank the audiences of these talks for their helpful remarks and suggestions. In particular, I would like to thank Alexandru Baltag, Johan van Benthem, Jan van Eijck, Jan Heylen, Emiliano Lorini, Alexandru Marcoci, Ahti-Veikko Pietarinen, Sonja Smets, Jean Paul Van Bendegem and two anonymous referees for their feedback on earlier versions of this paper. This research was supported by a PhD fellowship of the Research Foundation – Flanders (FWO).

[1] These philosophical debates will not be directly addressed in this paper; for overviews, the reader can consult [45] and [5], respectively.

doi: 10.2143/LEA.230.0.3141810.

This enterprise is motivated by a variety of interrelated issues. In the first place, a logical perspective on surprise can help to elucidate the basic properties of this notion. Starting from the concrete empirical results about surprise, a complete axiomatization is proposed in which the observed behavioral patterns can be derived as theorems. In other words, the fundamental laws of surprise can be 'reverse engineered' out of the concrete behavior that they generate. Secondly, the resulting logical system serves as a highly expressive language to formally specify agent architectures; it belongs to the general framework of (dynamic) epistemic logic, which is becoming a contemporary 'lingua franca' in multi-agent systems [52, 42]. Thirdly, and most importantly, this project constitutes a concrete illustration of the so-called *dynamic turn* in logic [48, 49]. According to this position, many theorems, phenomena, etc. which are usually expressed or analyzed in an entirely statical way, actually have a lot of dynamics going on, and could benefit significantly from analyses which explicitly represent this underlying dynamics. Several illustrations of this dynamic turn stem from the field of game theory [10, 11, 47]. Considering the main conceptual thesis of this paper (as stated above), it should be clear that the paper offers a new illustration of the dynamic turn in logic (coming from cognitive science, rather than game theory).

The remainder of the paper is organized as follows. Section 2 briefly reviews the literature on surprise in cognitive science, multi-agent systems and logic. In Section 3 I argue that two earlier formalizations do not adequately represent the dynamic nature of surprise, and make some suggestions on how this can be achieved. In Section 4, then, I show how these suggestions can be developed into a full-fledged dynamic logic of surprise, which can capture several key aspects of surprise, such as its transitory (short-lived) nature and its role in belief revision. Finally, Section 5 wraps things up, and discusses some potential lines of research which will be explored in future work.

2. Three Perspectives on Surprise

This section provides an overview of the literature on surprise in cognitive science, multi-agent systems, and logic, focusing on those topics and debates that are most relevant for our current purposes. For more comprehensive overviews, the reader can fruitfully consult [22, 28, 37].

2.1. *Cognitive Science*

The emotion of surprise is probably of old phylogenetic origin [36]. This short-lived state of mind is caused in an agent when she encounters an

event that she did not expect. Surprise comes in degrees of intensity, which depend monotonically on the degree of unexpectedness of the surprise-causing event [44]. Like most emotions, surprise has both phenomenal aspects (there is an experience of "what it is like to be surprised" [32, 38]) and physical (behavioral/physiological) manifestations, such as a characteristic facial expression (raised eyebrows, opened mouth, etc.) and a decrease in heart rate [31, 43].

The cognitive-psychoevolutionary theory of surprise [30] claims that typically, an unexpected event elicits a sequence of four processes. First, the event is appraised as unexpected, i.e. as conflicting with a previously held belief.[2] Secondly, if the degree of unexpectedness is sufficiently large, then ongoing processes are interrupted and attention is shifted to the unexpected event. Thirdly, the unexpected event is analyzed and evaluated, which can lead to the fourth process, viz. revision of the relevant beliefs.

The fact that this sequence ends in belief revision helps to explain the transitory (short-lived) character of surprise. When a surprising event occurs again and again, subjects tend to 'get used' to it, and after a few occurrences they do not find it surprising at all anymore [4, Experiment II]. Initially, the surprising event is unexpected: it conflicts with a previously held belief B. This leads to a process of belief revision, which removes B from the agent's stock of beliefs (and perhaps replaces it with another belief). When the same event happens again, it is no longer surprising, because it no longer conflicts with a previously held belief (in particular, it does not conflict with B anymore).

The third step in the sequence of events triggered by an unexpected event involves *analyzing* that event. One of the features that is typically analyzed, is the event's cause: does it have a 'substantial' cause, or should it be attributed to 'mere chance'? Surprise thus leads to 'causal curiosity': it motivates the agent to inquire about the event's cause [30, 44].[3] Charlesworth has compared the motivational power of *unexpected (surprising) data* (which conflict with a previously held belief), *expected data* (which are in full agreement with previously held beliefs) and *novel data* (about which the agent had no previous beliefs at all), and his experiments show that surprising data have the highest motivational power, i.e. they trigger further inquiry most frequently [4].

I just mentioned Charlesworth's distinction between *unexpected* and *novel* data. For an event to be unexpected, it really has to conflict with a

[2] Building upon earlier work on schema theory [39, 41], the cognitive-pyschoevolutionary theory uses the notion of 'schema' rather than 'belief'. This distinction is not relevant for our current purposes, so I will simply use the term 'belief'.

[3] The process of searching for an explanation of an observed event is widely known as *abduction*. Peirce, who coined this term, explicitly refers to surprising events when characterizing abductive reasoning [34, paragraph 189].

previously held belief; if the agent did not have any beliefs about that (type of) event(s), then the event is not unexpected, but rather novel. The most common perspective is that surprise can only be generated by unexpected data, not by novel data [4, 30, 44].[4] However, some theorists maintain that an agent can also be surprised about events that she previously did not have any beliefs about. For example, Ortony and Partridge [33] distinguish between *actively expected* events and *passively expected* or *assumed* events, and claim that surprise can arise from active expectation failure as well as assumption failure.[5] There is no real contradiction between both perspectives, since Ortony and Partridge maintain that in the case of surprise caused by assumption failure, the agent still has a belief, albeit a 'passive' one (an assumption). For example, if the legs of my chair suddenly break and I fall, I am surprised, not because I actively believed that I would remain seated in the chair, but because I passively expected (i.e. assumed) that the chair's legs are strong enough to support me. The tension between both perspectives can thus be resolved by postulating implicit beliefs with which the novel event conflicts—hence, although the event does not conflict with any explicit (active) beliefs, it *does* conflict with the postulated implicit (passive) belief.[6]

2.2. *Multi-Agent Systems*

Since surprise typically leads to processes of learning and belief revision in humans, it is a natural move to endow *artificial agents* with the capability of feeling surprise, which can guide them in their actions. In a recent series of papers, Macedo and Cardoso have done exactly this [25, 26, 27, 23, 24]. This work is based on the cognitive theories of surprise described in the previous subsection [30, 33], and can thus also be seen as a simulation of the human surprise mechanism (with various simplifications, obviously).

The agent's goal is to explore an unknown and dynamic environment. The agent architecture is similar to the BDI (belief-desire-intention) architecture [52], and looks as follows. The agent's *perceptual system* provides

[4] This perspective is also common among philosophers. Davidson, for example, claims that "I could not be surprised […] if I did not have beliefs in the first place. […] Surprise requires that I be aware of a contrast between what I did believe and what I come to believe" [6, p. 326].

[5] Peirce, too, claims that surprise "has its Active and its Passive variety;—the former when what one perceives positively *conflicts* with expectation, the latter when having no positive expectation but only the absence of any suspicion of anything out of the common something quite unexpected occurs" [35, paragraph 315].

[6] For example, novel events "can also be conceptualised as instances of expectancy disconfirmation: They disconfirm the *implicit*, schema-based belief that the unexpected event was unlikely to occur in the given situation." [44, p. 6, my emphasis].

(partial) information about the environment, and stores it in *memory*. When new (hypothetical) information comes in, the agent's *surprise-generating module* calculates the intensity of the surprise caused by that piece of information. Finally, the *decision-making module* selects the agent's next action by considering, for every available action α, how surprised the agent would be by the state of the world caused by α, and then selecting the action that maximizes the agent's anticipated surprise. This module thus implements a utility-maximizing function, where the agent's utility is assumed to coincide with her anticipated surprise (more sophisticated architectures also take other emotions into account).

In the simplest model [26], the anticipated intensity of surprise elicited by a piece of information φ is calculated as follows:[7]

$$S(\varphi) := 1 - P(\varphi). \tag{1}$$

The unexpectedness of φ is represented by $1-P(\varphi)$. Here, $P(\varphi)$ denotes the subjective probability of φ, which is computed based on frequencies stored in the agent's memory. Thus (1) clearly shows that the intensity of surprise about φ is a monotone increasing function of the unexpectedness of φ (cf. supra).

This work on surprise-based agent architectures fits in the broader field of emotion-based agent architectures [2, 14, 22]. There are also proposals to incorporate the *dynamics* of emotion [3, 13, 29], but none of them so far make use of the framework of (probabilistic) dynamic epistemic logic.

2.3. *Logic*

Lorini has argued that researchers attempting to incorporate surprise and other emotions into multi-agent systems can benefit from the accuracy of logical frameworks for the formal specification of emotions [21]. Together with Castelfranchi, he has developed a logical framework for surprise [19, 20]. Just like Macedo and Cardoso's, this framework is based on the cognitive theories of surprise described in Subsection 2.1 [30, 33], and can thus be seen as a formal-logical model of human surprise.

I will now discuss the main features of this framework.[8] The base logic is a system of probabilistic epistemic logic with a belief operator B and

[7] There exist more complex (and realistic) proposals for defining surprise in terms of unexpectedness (probability) [23]. However, the experimental data do not seem to single out one of these complex definitions over the other ones. Furthermore, the main conceptual points of this paper (regarding the dynamic nature of surprise) can perfectly be made using (1). Therefore, I will stick to the simpler definition.

[8] In this subsection in particular, I will not be able to do justice to all details of the framework under discussion. For example, I will only reason 'within' the logic, and not say

formulas about (linear combinations of) probabilities, such as $P(\varphi) \geq 0.5$ and $P(\varphi) + 2P(\psi) \geq 0.7$ [15]. This system is extended with PDL-style dynamic operators [17], and two unary operators *Test* and *Datum*. The formulas $Test(\varphi)$ and $Datum(\varphi)$ are to be read as "the agent is currently scrutinizing φ" and "the agent has perceptual datum φ", respectively. Furthermore, there are actions $observe(\varphi)$ and $retrieve(\varphi)$, which represent observing that φ is the case and retrieving (from memory) that φ. Each of these actions give rise to a PDL-style dynamic operator. The two most important axioms are:

$$[observe(\varphi)]\, Datum(\varphi), \tag{2}$$

$$[retrieve(\varphi)]\, Test(\varphi). \tag{3}$$

Axiom (2) says that after the agent observes that φ, this becomes a perceptual datum; analogously, axiom (3) says that after the agent has retrieved φ, this becomes an item under scrutiny.

With these resources, the notion of *mismatch-based surprise* can be defined. This emotion arises when there is a conflict between a perceptual datum ψ and a currently scrutinized belief φ; 'conflict' here means that the agent believes that φ and ψ cannot be jointly true. Furthermore, the *intensity* of a mismatch-based surprise is defined as the probability that the agent assigns to the scrutinized belief φ. Hence, the more confident the agent is in her belief that φ, the more intensely she will be surprised upon receiving a perceptual datum that conflicts with φ (this captures exactly the idea that the intensity of surprise is a monotone function of the degree of unexpectedness). Formally:

$$MismatchS(\psi,\varphi) \;:\equiv\; Datum(\psi) \wedge Test(\varphi) \wedge B(\psi \rightarrow \neg\varphi), \tag{4}$$

$$IntensityS(\psi,\varphi) = c \;:\equiv\; MismatchS(\psi,\varphi) \wedge P(\varphi) = c. \tag{5}$$

3. Surprise as a Dynamic Phenomenon

In this section I will argue that neither Macedo and Cardoso's computational nor Lorini and Castelfranchi's logical models of surprise adequately capture the dynamic nature of surprise. Afterwards I will suggest how the dynamics of surprise *can* adequately be formalized.

anything about its formal semantics. Furthermore, Lorini and Castelfranchi define *two* types of surprise, mismatch-based surprise and astonishment, but I will only discuss the first one, because it is better suited to illustrate the main claims of the next section. (However, one might argue that the notion of surprise defined in Section 4 is actually closer to astonishment than to mismatch-based surprise.)

3.1. *Quasi-Static Analyses of Surprise*

Let's first fix some terminology. Surprise is caused by an unexpected event. Any mental state (beliefs, desires, emotions, etc.) that the agent had (just) *before* perceiving the unexpected event will be called *'prior'*; any such state that she has (just) *after* perceiving the event will be called *'posterior'*.[9] A statement that involves only prior notions or only posterior notions will be called 'temporally coherent'; a statement that involves both prior and posterior notions will be called 'temporally incoherent'.

Consider Macedo and Cardoso's analysis of surprise, and recall their Definition (1) of surprise intensity as unexpectedness:

$$S(\varphi) = 1 - P(\varphi).$$

The left side contains a posterior notion: the intensity of the surprise felt by the agent after the unexpected event. The right side, however, contains a prior notion: the agent's subjective probability before the unexpected event. Hence, Definition (1) is a temporally incoherent statement.

To see this more clearly, note that there are two ways of reading (1) as a temporally coherent statement: (i) by considering both S and P to be prior notions, and (ii) by considering both S and P to be posterior notions. For interpretation (i), consider a case where the agent assigns a low (prior) probability to φ; Definition (1) then says that she should experience a highly intensive surprise about φ. Under interpretation (i), this surprise is prior; in other words, the agent is highly surprised about an event *before* she has even perceived it—which is clearly absurd. For interpretation (ii), consider a case where the agent is highly surprised after perceiving an occurrence of φ; Definition (1) then says that she assigns a low probability to φ. Under interpretation (ii), this probability is posterior; in other words, even after the agent has observed an occurrence of φ, she still assigns a low probability to it—which clearly contradicts the common assumption that agents process new information via Bayesian updating.[10]

I now turn to Lorini and Castelfranchi's analysis of surprise. Let's first consider the qualitative notion of mismatch-based surprise—ignoring, for the moment, surprise intensity. Recall their Definition (4):

$$MismatchS(\psi,\varphi) \equiv Datum(\psi) \wedge Test(\varphi) \wedge B(\psi \rightarrow \neg\varphi).$$

[9] This terminology is analogous to the use of 'priors' and 'posteriors' in Bayesian frameworks. However, it should be emphasized that in this paper, 'prior' and 'posterior' are defined in terms of (being before or after) *perceiving* the unexpected event, while in Bayesian frameworks they are defined in terms of (being before or after) the *probabilistic update* ('probability revision') triggered by that event.

[10] And $P(\varphi|\varphi) = 1$, so after the occurrence of φ, the agent should assign probability 1 to it.

The left side contains a posterior notion: the agent's mismatch-based surprise after the unexpected event. The right side is more complicated. The first conjunct is posterior: ψ is only a perceptual datum after it has been observed by the agent; this dynamics was explicitly represented in (2). The second conjunct is both prior *and* posterior: φ was under scrutiny before the observation of the unexpected event, and remains so afterwards. The third and final conjunct is prior: the agent believed that ψ and φ cannot be jointly true before the unexpected event; typically, she will drop this belief as a result of her surprise (recall from Subsection 2.1 that surprise typically leads to a process of belief revision). Thus, in total, Definition (4) is temporally incoherent.[11]

Finally, let's consider the quantitative aspects of Lorini and Castelfranchi's system. Recall their Definition (5) of surprise intensity:

$$IntensityS(\psi,\varphi) = c \equiv MismatchS(\psi,\varphi) \wedge P(\varphi) = c.$$

The left side contains a posterior notion: the intensity of the agent's mismatch-based surprise after she has perceived the unexpected event. The right side is, again, more complicated. The first conjunct—which was also the left side of (4)—is posterior: the agent experiences mismatch-based surprise only after perceiving the unexpected event. The second conjunct, however, involves a prior notion, viz. the probability that the agent assigns to the scrutinized item φ before perceiving the unexpected event. Hence, Definition (5) is temporally incoherent as well.[12]

An intuitively right principle about surprise should look somewhat like this: if the agent has a (prior) belief that ψ and φ are incompatible, and assigns (prior) probability c to φ, then after retrieving φ and observing an occurrence of ψ, she will experience a (posterior) mismatch-based surprise with intensity c. Formally, this looks as follows:[13]

$$(B(\psi \rightarrow \neg\varphi) \wedge P(\varphi) = c) \rightarrow [retrieve(\varphi); observe(\psi)]IntensityS(\psi,\varphi) = c. \tag{6}$$

[11] Again, there are two ways of reading (4) as a temporally coherent statement: by considering all notions that appear in it to be prior, or by considering all those notions to be posterior. It is easy to see, however, that both interpretations quickly lead to counterintuitive consequences. Similar remarks apply to (5), which will be discussed next.

[12] It should be emphasized that the assessment of Lorini and Castelfranchi's analysis as temporally incoherent is only valid on the assumption that the terms 'prior' and 'posterior' are defined relative to the moment of *perceiving* the unexpected event, as specified at the beginning of this subsection (also recall Footnote 3.1). In particular, if these terms are defined relative to the moment of *recognizing the mismatch* between the datum and the scrutinized expectation—which is the viewpoint taken by Lorini and Castelfranchi themselves—, then this analysis *is* temporally coherent. Thanks to an anonymous referee for some helpful discussion about this.

[13] As usual, ';' denotes ordinary PDL sequential composition [17]; this operation on actions is allowed in Lorini and Castelfranchi's system.

However, to derive (6) in Castelfranchi and Lorini's system, one needs principles such as (7) and (8), which link the agent's prior and posterior states by claiming that her observation of the occurrence of ψ does not change her relevant beliefs and probabilities in any way. This is highly counterintuitive: both (7) and (8) run entirely against the idea that surprise triggers a process of belief revision; additionally, (8) clearly contradicts the common assumption that agents process new information via Bayesian conditionalization.

$$B(\psi \to \neg\varphi) \;\to\; [observe(\psi)]\,B(\psi \to \neg\varphi), \tag{7}$$

$$P(\varphi) = c \;\to\; [observe(\psi)]\,P(\varphi) = c. \tag{8}$$

3.2. *Towards a Fully Dynamic Analysis of Surprise*

I have shown that both Macedo and Cardoso's definition of surprise intensity (1) and Lorini and Castelfranchi's definitions of mismatch-based surprise and its intensity (4–5) are temporally incoherent (but recall Footnote 3.1). There is a uniform explanation for this: surprise is an essentially dynamic phenomenon, but none of these authors explicitly represents this dynamics, so they have to 'smuggle' it into their systems—which thus end up being temporally incoherent.[14]

Before moving on, we need to clarify the relationship between the systems discussed above and the system that will be developed in this paper. The problem of temporal incoherence is situated at the *conceptual*, rather than at the *empirical* level, and is therefore largely independent of the original motivations behind the systems discussed above. For example, Lorini and Castelfranchi's goal is first and foremost to propose a cognitively realistic model of surprise; although their analysis is largely static, it certainly achieves its main goal, since it is highly successful at capturing various experimentally observed properties of surprise. In contrast, the main motivation behind this paper is to propose a temporally coherent model of surprise (using the framework of dynamic epistemic logic). Looking ahead, this means that the major advantage of the new account of surprise that will be developed in Section 4 will not be so much its level of empirical adequacy—which, I will argue, is more or less comparable to those of the other accounts—, but rather the fact that it is temporally coherent, and thus better able to capture the dynamic nature of surprise.

Now that this methodological issue has been clarified, we are ready to move on. To obtain a temporally coherent definition of surprise, which respects the different 'stages' (before vs. after perceiving the unexpected

[14] A similar story can be told about the role of dynamics in Aumann's celebrated 'agreeing to disagree' theorem in game theory [10, 11].

event), the dynamics of surprise needs to be represented explicitly. I will use a public announcement operator $[!\varphi]$ for this purpose (technical details will be discussed in the next section). Whether a certain notion is to be interpreted as prior or as posterior, is now encoded directly in the syntax of the language: if the notion is within the scope of a dynamic operator, it is posterior, otherwise it is prior. For example, $P(\varphi) = 0.2$ means that the agent's *prior* probability of φ is 0.2, while $[!\varphi]P(\varphi) = 0.2$ means that her *posterior* probability of φ is 0.2.

We will work with a simple measure of surprise intensity S, based on Macedo and Cardoso's (1).[15] When the surprise dynamics is explicitly represented, (1) is transformed into the following:

$$[!\varphi]S(\varphi) = c \leftrightarrow P(\varphi) = 1 - c. \tag{9}$$

This principle says that the agent assigns probability $1-c$ to φ before the unexpected event iff after that event she is surprised with intensity c. It thus says exactly the same as (1), but now in a temporally coherent way: both sides of (9) are prior statements.[16] Furthermore, note that the right-to-left direction of (9) is similar in spirit to (6), which was very intuitive, but which was only derivable using additional implausible principles such as (7–8).

4. Modeling Surprise in Probabilistic DEL

In the previous section, I made some suggestions on how the dynamics of surprise can be represented explicitly. In this section, these suggestions will be developed into a full-fledged logical system. I will also show how this system can naturally capture several important properties of surprise.

4.1. *The Logical System*

Given the dynamic nature of surprise, and its connection with epistemic states and processes (beliefs, unexpectedness, belief revision, etc.), it is natural to work in the general framework of dynamic epistemic logic. This framework is rapidly becoming a 'lingua franca' or 'universal toolbox', which has been applied to problems in game theory, philosophy, artificial intelligence, etc. [10, 8, 9, 15, 18, 50].

[15] Recall Footnote 2.2.

[16] The left formula *as a whole* is prior; the *subformula* $S(\varphi) = c$ occurs inside the scope of the $[!\varphi]$-operator, and is thus posterior. In other words, principle (9) is able to express a connection between the agent's *prior* probability and her *posterior* surprise intensity in a temporally coherent way, by making use of the dynamic $[!\varphi]$-operator.

Fix a countable set *Prop* of proposition letters. In this paper I will only work with a single agent, so it is not necessary to introduce agent indices. The formal language $\mathcal{L}$ is given by the following Backus-Naur form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K\varphi \mid \sum c_i P(\varphi) \geq c \mid S(\varphi) \geq c \mid S(\varphi) \leq c \mid [!\varphi]\varphi,$$

where $p \in Prop$ and $c_i, c \in \mathbb{Q}$. As usual, $K\varphi$ means that the agent knows that φ. Similarly, $P(\varphi) \geq c$ means that the agent assigns probability (degree of belief) at least c to φ. Arbitrary linear combinations of probability terms are allowed mainly for technical reasons that need not concern us here [15]. Because of this generality, any type of (in)equality of probabilities is expressible [18, Def. 2]. The formula $S(\varphi) \geq c$ says that the agent is surprised about φ with intensity at least c. Here, full expressivity is not allowed, and so the $\geq$- and $\leq$-forms are both taken as primitive. One can then define $S(\varphi) < c$ as $\neg(S(\varphi) \geq c)$, etc.

Finally, $[!\varphi]\psi$ should be read as: "after a public announcement of φ, it will be the case that ψ". Its dual is $\langle !\varphi \rangle \psi := \neg[!\varphi]\neg\psi$. Public announcement is usually explicated in terms of rational communication, but actually, almost any public event can be modeled using public announcements; for example, a strike of lightning can be modeled as a public announcement of the proposition 'lightning occurs (at time t and location ℓ)'.[17] It thus makes perfect sense to represent an unexpected event (whatever its exact nature) as a public announcement.[18]

We now turn to the models on which this language will be interpreted:

Definition 1. A *surprise model* is a tuple $\mathbb{M} := \langle W, R, \mu, \sigma, V \rangle$, where W is a non-empty and finite set of states, R is an equivalence relation on W, and $V : Prop \to \wp(W)$ is a valuation function. Furthermore, μ assigns to every state $w \in W$ a probability mass function $\mu(w) : W \to [0,1]$ that satisfies two conditions: (i) if $(w,v) \notin R$ then $\mu(w)(v) = 0$, and (ii) $\mu(w)(w) > 0$. Finally, σ assigns to every state $w \in W$ a surprise measure, i.e. a partial function $\sigma(w) : \wp(W) \rightharpoonup [0,1]$.

[17] Van Benthem, Gerbrandy and Kooi make a similar comment: "While much of the theory has been developed with conversation and communication in mind, it is important […] to stress that we are not doing some sort of formal linguistics. The formal systems we will be dealing with apply just as well to observation, experimentation, learning, or any sort of information-carrying scenario." [46, p. 71].

[18] This also resolves a terminological tension in the literature on surprise. Agents are surprised *about* some *propositional* content (a piece of information), but their surprise is *caused by* some (non-propositional) *event*. In the new system, the propositional content of the surprise is formalized as the proposition φ, while its cause is formalized as the *public announcement of* that proposition. In short: φ is a proposition, but $!\varphi$ is an event.

Definition 2. The class of all surprise models will be denoted $\mathcal{C}_S$. Furthermore, $\mathcal{C}_S^*$ is the class of all surprise models whose surprise measures are entirely undefined, i.e. such that $\sigma(w)(X)$ is undefined for all $w \in W$ and $X \subseteq W$.

A surprise model is thus just an ordinary finite[19] Kripke model $\langle W, R, V \rangle$ with additional components μ and σ. First of all, $\mu(w)(v) = c$ means that at state w, the agent assigns probability c to v being the actual state. Similarly, $\sigma(w)(X) = c$ means that at state w, the agent experiences surprise with intensity c about X (i.e. about one of the states in X being the actual state). Note the following differences between $\mu(w)$ and $\sigma(w)$ (for any state $w \in W$):

- $\mu(w)$ is a total function, so $\mu(w)(v)$ is defined for every state $v \in W$; on the other hand, $\sigma(w)$ is a partial function, so it is allowed that $\sigma(w)(X)$ is undefined for some sets $X \subseteq W$,
- $\mu(w)$ is required to satisfy conditions (i) and (ii), whose motivation is discussed extensively in [10, 11, 8, 9]; on the other hand, $\sigma(w)$ is not required to satisfy any additional conditions whatsoever,
- $\mu(w)$ is defined on individual states, and can additively be lifted to sets of states: $\mu(w)(X) = \sum_{x \in X} \mu(w)(x)$ (this essentially reflects the finite additivity of probabilities); on the other hand, $\sigma(w)$ is defined directly on sets of states, so it might happen that $\sigma(w)(\{x, y\}) \neq \sigma(w)(\{x\}) + \sigma(w)(\{y\})$.

These differences show that unlike the well-behaved epistemological notion of probability (degree of belief), the psychological notion of (degree of) surprise satisfies no static regularities whatsoever. This is a clear manifestion of the essentially dynamic nature of surprise in the definition of surprise models.[20]

I now turn to the logic's semantics. This is entirely as expected; the formal clauses are stated in Definition 3. Given a formula $\varphi \in \mathcal{L}$ and a surprise model $\mathbb{M}$, I use $[\![\varphi]\!]^{\mathbb{M}}$ to denote the set $\{w \in W \mid \mathbb{M}, w \models \varphi\}$. The clause

[19] The assumption that surprise models are finite ensures that probabilities can be represented using simple probability mass functions. This assumption can be dropped; the general case uses σ-algebras to represent probabilities [9, 40]. However, the main points of this paper are of a more conceptual nature, and can perfectly be made using the less sophisticated setup.

[20] One might consider adding the requirements that if $X \subseteq Y \subseteq W$, then $\sigma(w)(X) \geq \sigma(w)(Y)$ and $\sigma(w)(W - X) = 1 - \sigma(w)(X)$, in analogy to the well-known Kolmogorov axioms for probability. However, the only motivation for such requirements seems to be the observation that "surprise is inversely correlated with probability", which is only plausible if 'surprise' is read as posterior and 'probability' as prior. I will return to this suggestion after the dynamics has been formally introduced (cf. Lemma 9). Thanks to an anonymous referee for some useful discussion about this.

for surprise formulas holds for $\gtrless \in \{\geq, \leq\}$; I will return to it later (see Lemma 7). Note that to interpret a formula of the form $[!\varphi]\psi$ at a surprise model $\mathbb{M}$, the subformula ψ has to be interpreted at the updated model $\mathbb{M}\restriction\varphi$, which is well-defined because of Definition 4 and Lemma 6. Finally, Definition 5 states the usual definition of semantic validity.

Definition 3. Consider a surprise model $\mathbb{M}$ and state w in $\mathbb{M}$. Then:

$$
\begin{array}{lll}
\mathbb{M}, w \models p & \text{iff} & w \in V(p) \qquad\qquad (\text{for } p \in Prop), \\
\mathbb{M}, w \models \neg\varphi & \text{iff} & \mathbb{M}, w \not\models \varphi, \\
\mathbb{M}, w \models \varphi \wedge \psi & \text{iff} & \mathbb{M}, w \models \varphi \text{ and } \mathbb{M}, w \models \psi, \\
\mathbb{M}, w \models K\varphi & \text{iff} & \text{for all } v \in W: \text{ if } wRv \text{ then } \mathbb{M}, v \models \varphi, \\
\mathbb{M}, w \models \sum c_i P(\varphi_i) \geq c & \text{iff} & \sum c_i \mu(w)([\![\varphi_i]\!]^{\mathbb{M}}) \geq c, \\
\mathbb{M}, w \models S(\varphi) \gtrless c & \text{iff} & \begin{cases} \sigma(w)([\![\varphi]\!]^{\mathbb{M}}) \gtrless c & \text{if } \sigma(w)([\![\varphi]\!]^{\mathbb{M}}) \text{ is defined} \\ c = 0 & \text{otherwise,} \end{cases} \\
\mathbb{M}, w \models [!\varphi]\psi & \text{iff} & \text{if } \mathbb{M}, w \models \varphi \text{ then } \mathbb{M}\restriction\varphi, w \models \psi.
\end{array}
$$

Definition 4. Consider an arbitrary surprise model $\mathbb{M} = \langle W, R, \mu, \sigma, V\rangle$ and formula $\varphi \in \mathcal{L}$, and suppose that $\mathbb{M}, w \models \varphi$ for some $w \in W$. Then the updated model $\mathbb{M}\restriction\varphi := \langle W^{\varphi}, R^{\varphi}, \mu^{\varphi}, \sigma^{\varphi}, V^{\varphi}\rangle$ is defined as follows:

- $W^{\varphi} := [\![\varphi]\!]^{\mathbb{M}} = \{w \in W \mid \mathbb{M}, w \models \varphi\}$,
- $R^{\varphi} := R \cap ([\![\varphi]\!]^{\mathbb{M}} \times [\![\varphi]\!]^{\mathbb{M}})$,
- $\mu^{\varphi}(w)(v) := \frac{\mu(w)(v)}{\mu(w)([\![\varphi]\!]^{\mathbb{M}})}$ for all $w, v \in W^{\varphi}$,
- $\sigma^{\varphi}(w)(X) := 1 - \mu(w)(X)$ for all $w \in W^{\varphi}$, $X \subseteq W^{\varphi}$,
- $V^{\varphi}(p) := V(p) \cap [\![\varphi]\!]^{\mathbb{M}}$ for every $p \in Prop$.

Definition 5. For any formula $\varphi \in \mathcal{L}$ and class of models $\mathcal{C}$, we say that $\mathcal{C} \models \varphi$ iff $\mathbb{M}, w \models \varphi$ for all models $\mathbb{M} \in \mathcal{C}_{\mathcal{S}}$ and states w in $\mathbb{M}$.

Lemma 6. *The class $\mathcal{C}_{\mathcal{S}}$ is closed under public announcements, i.e. if $\mathbb{M} \in \mathcal{C}_{\mathcal{S}}$, then also $\mathbb{M}\restriction\varphi \in \mathcal{C}_{\mathcal{S}}$ (for any formula $\varphi \in \mathcal{L}$). This does not hold for $\mathcal{C}^*_{\mathcal{S}}$.*

Proof. The $\mathcal{C}_{\mathcal{S}}$ case is trivial: for the non-surprise components, see [10, Lemma 9], and since Definition 1 does not require the surprise measures to satisfy any additional requirements, there is nothing else to prove. For $\mathcal{C}^*_{\mathcal{S}}$, note that by Definition 4, the updated surprise measures are total functions, even if the original surprise measures were entirely undefined. □

The public announcement of φ in a model $\mathbb{M}$ deletes all $\neg\varphi$-states from that model; this is a standard idea [50]. The probability functions are changed by Bayesian conditionalization on the announced proposition φ [8, 10, 9, 18]. To see this more clearly, note that the definition of the updated probability function can be rewritten using conditional probabilities: $\mu^{\varphi}(w)(v) = \mu(w)(v|[\![\varphi]\!]^{\mathbb{M}})$. Most importantly, the updated surprise measure $\sigma^{\varphi}(w)$ is defined in terms of the original probability function $\mu(w)$. This is the only substantial property of surprise that is assumed in the logic's semantic setup; it is clearly of a dynamic nature (linking the original and the updated model).

Even though the surprise measures $\sigma(w)$ are allowed to be partial, Lemma 7 below shows that this does not lead to any truth value gaps in the semantics. When we are modeling concrete scenarios, we typically want to assume that the agent initially (i.e. before any unexpected events have taken place) experiences no surprise. Lemma 7 therefore justifies the following heuristic rule (HEUR):

> When modeling a scenario, it can be assumed that the 'initial' model $\mathbb{M}$ (which represents the situation before any unexpected events have taken place) leaves all surprise measures undefined, i.e. that $\mathbb{M} \in \mathcal{C}_S^*$.

Lemma 7. *Consider an arbitrary surprise model* $\mathbb{M} = \langle W, R, \mu, \sigma, V\rangle$ *and formula* $\varphi \in \mathcal{L}$*, and suppose that* $\sigma(w)([\![\varphi]\!]^{\mathbb{M}})$ *is undefined. Then* $\mathbb{M}, w \vDash S(\varphi) = 0$.

Proof. Since $\sigma(w)([\![\varphi]\!]^{\mathbb{M}})$ is undefined, it follows by the semantic clause for $S(\varphi) \geq c$ that $\mathbb{M}, w \vDash S(\varphi) \geq 0$ (and $\mathbb{M}, w \nvDash S(\varphi) \geq c$ for all $c \neq 0$). Entirely analogously, $\mathbb{M}, w \vDash S(\varphi) \leq 0$ (and $\mathbb{M}, w \nvDash S(\varphi) \leq c$ for all $c \neq 0$). □

The following lemma states that the language $\mathcal{L}$ contains no redundancies. In particular, the surprise operator cannot be defined in terms of the other available operators.

Lemma 8. *There exists no formula* $\varphi \in \mathcal{L} - \{S\}$ *such that* $\vDash \varphi \leftrightarrow S(p) \geq 0.5$.

Proof. Consider the surprise models $\mathbb{M}_1$ and $\mathbb{M}_2$, defined as follows:

- $\mathbb{M}_1 = \langle W_1, R_1, \mu_1, \sigma_1, V_1\rangle$, $W_1 = \{w_1\}$, $R_1 = \{(w_1, w_1)\}$, $\mu(w_1)(w_1) = 1$, $\sigma_1(w_1)(X) = 0.6$ for all $X \subseteq W_1$, and $V_1(p) = W_1$,
- $\mathbb{M}_2 = \langle W_2, R_2, \mu_2, \sigma_2, V_2\rangle$, $W_2 = \{w_2\}$, $R_2 = \{(w_2, w_2)\}$, $\mu_2(w_2)(w_2) = 1$, $\sigma_2(w_2)(X) = 0.4$ for all $X \subseteq W_2$, and $V_2(p) = W_2$.

One can show by induction on the complexity of φ that

$$\text{for all } \varphi \in \mathcal{L} - \{S\} : \mathbb{M}_1, w_1 \vDash \varphi \text{ iff } \mathbb{M}_2, w_2 \vDash \varphi.$$

But it also holds that $\mathbb{M}_1, w_1 \vDash S(p) \geq 0.5$, while $\mathbb{M}_2, w_2 \nvDash S(p) \geq 0.5$. □

The distinction between the original and the updated model corresponds exactly to the distinction between prior and posterior notions that was introduced in the previous section. Hence, the definition $\sigma^{\varphi}(w)(X) = 1 - \mu(w)(X)$ defines *posterior* surprise in terms of *prior* probability. As a consequence, all the properties of probability are manifested in the posterior surprise measure (recall Footnote 4.1):

Lemma 9. *Consider an arbitrary surprise model* $\mathbb{M} = \langle W, R, \mu, \sigma, V \rangle$ *and formula* $\varphi \in \mathcal{L}$*, and suppose that* $\mathbb{M}, w \vDash \varphi$ *for some* $w \in W$*. For all* $w \in W^{\varphi}$ *and* $X \subseteq Y \subseteq W^{\varphi}$*, it holds that* $\sigma^{\varphi}(w)(X) \geq \sigma^{\varphi}(w)(Y)$ *and that* $\sigma^{\varphi}(w)(W - X) = 1 - \sigma^{\varphi}(w)(X)$.

Proof. Both items follow immediately from the definition of σ^{φ} and the fact that $\mu(w)$ is a probability mass function. For example, if $X \subseteq Y$, then $\mu(w)(X) \leq \mu(w)(Y)$, and hence $\sigma^{\varphi}(w)(X) = 1 - \mu(w)(X) \geq 1 - \mu(w)(Y) = \sigma^{\varphi}(w)(Y)$. □

Before moving to the logic's proof theory, I will illustrate and justify its semantics by discussing a simple example in full detail.

Example 10. Consider the following scenario. Mary does not know whether it is currently snowing. In fact, it is indeed currently snowing, but since Mary does not yet know about this, she experiences no surprise about it whatsoever. Furthermore, since it is July and Mary knows that snow in July is very rare at her current location, she considers it very unlikely that it is currently snowing. This example can be formalized using the following surprise model: $\mathbb{M} = \langle W, R, \mu, \sigma, V \rangle$, $W = \{w, v\}$, $R = W \times W$, $\mu(w)(w) = \mu(v)(w) = 0.05$, $\mu(w)(v) = \mu(v)(v) = 0.95$, $V(p) = \{w\}$, and $\sigma(w)(X)$ and $\sigma(v)(X)$ undefined for all $X \subseteq W$. (Note that we have followed the heuristic rule Heur discussed above.) The proposition letter p represents 'it is snowing'; the state w represents the actual world. This model is a faithful representation of the scenario described above; for example:

$$\mathbb{M}, w \vDash \neg Kp \wedge \neg K\neg p \wedge P(p) = 0.05 \wedge P(\neg p) = 0.95 \wedge S(p) = 0.$$

Now suppose that Mary goes outside and sees that it is actually snowing. This can be modeled as a public announcement of p (recall Footnote 4.1). Applying Definition 4, we obtain the updated model $\mathbb{M} \restriction p$, with $W^p = \{w\}$, $R = \{(w, w)\}$,

$$\mu^p(w)([\![p]\!]^{\mathbb{M}\restriction p}) = \mu^p(w)(w) = \frac{\mu(w)(w)}{\mu(w)([\![p]\!]^{\mathbb{M}})} = \frac{\mu(w)(w)}{\mu(w)(w)} = 1,$$

$$\sigma^p(w)([\![p]\!]^{\mathbb{M}\restriction p}) = \sigma^p(w)(\{w\}) = 1 - \mu(w)(\{w\}) = 1 - 0.05 = 0.95.$$

Using this updated model $\mathbb{M} \upharpoonright p$, we find that

$$\mathbb{M}, w \models [!p](Kp \wedge P(p) = 1 \wedge P(\neg p) = 0 \wedge S(p) = 0.95).$$

So after going outside, Mary comes to know that it is in fact snowing. She also adjusts her probabilities: she is now certain that it is snowing; i.e. she assigns probability 1 to p being true and probability 0 to p being false. These are the main *cognitive* effects of Mary's observation that it is snowing. However, on the *emotional* side, she is also highly surprised to find out that it is snowing, because she initially considered this highly unlikely. These are the results that one would intuitively expect, so the semantic setup introduced above seems to yield an adequate representation of (the interactions between) the cognitive (epistemic and probabilistic) and emotional (surprise) effects of a public announcement.

I now turn to the logic's proof theory. *Reduction axioms* are equivalences which allow us to push the public announcement operator through any of the other connectives, thus yielding an effective procedure to rewrite any formula as an equivalent formula that doesn't contain any dynamic operators. The reduction axioms for all operators of $\mathcal{L} - \{S\}$ are well-known [10, 11, 18, 50], cf. items 1-5 of Definition 11 below. What about reduction axioms for S? Recall that in Subsection 3.2 I suggested a dynamified (and temporally coherent!) version (9) of Macedo and Cardoso's original (1). With only minor modifications,[21] this suggestion can be turned into reduction axioms for S; cf. items 6-7 below.

Definition 11. The reduction axioms for public announcement:

1. $[!\varphi]p \leftrightarrow \varphi \rightarrow p$ (for $p \in Prop$)
2. $[!\varphi]\neg\psi \leftrightarrow \varphi \rightarrow \neg[!\varphi]\psi$
3. $[!\varphi](\psi_1 \wedge \psi_2) \leftrightarrow [!\varphi]\psi_1 \wedge [!\varphi]\psi_2$
4. $[!\varphi]K\psi \leftrightarrow \varphi \rightarrow K[!\varphi]\psi$
5. $[!\varphi]\sum c_i P(\psi_i) \geq c \leftrightarrow \varphi \rightarrow \sum c_i(\langle !\varphi\rangle\psi) \geq cP(\varphi)$
6. $[!\varphi]\, S(\psi) \geq c \leftrightarrow \varphi \rightarrow P(\langle !\varphi\rangle\psi) \leq 1 - c$
7. $[!\varphi]\, S(\psi) \leq c \leftrightarrow \varphi \rightarrow P(\langle !\varphi\rangle\psi) \geq 1 - c$

We are now ready to axiomatize the logic of surprise.

[21] Trivial modifications are that the statement about = needs to be 'split out' into statements about $\leq$ and $\geq$, and that in the reduction axioms the argument of S should be an arbitrary formula ψ, and not just φ itself. A more serious modification is that the right sides of the reduction axioms should not contain simply $P(\psi)$, but rather $P(\langle !\varphi\rangle\psi)$, to 'pre-encode' the effect of the public announcement of φ on ψ.

Definition 12. $\mathcal{SURPRISE}$ *i*s the logic axiomatized as follows:

- all of propositional logic,
- S5 for the knowledge operator K,
- the Kolmogorov axioms for the probability operator P [10, 8, 7, 9, 15, 18]:
  - $P(\varphi) \geq 0$,
  - $P(\top) = 1$,
  - $P(\varphi \wedge \psi) + P(\varphi \wedge \neg\psi) = P(\varphi)$,
  - if $\vdash \varphi \leftrightarrow \psi$ then $\vdash P(\varphi) = P(\psi)$
- some auxiliary axioms for linear inequalities (described in [8, 9, 15, 18]),
- the axioms $K\varphi \to P(\varphi) = 1$ and $\varphi \to P(\varphi) > 0$ (these correspond to conditions (i) and (ii) in Definition 1 [10, 11, 8, 9]),
- some auxiliary axioms and rules for the surprise operator S:
  - $S(\varphi) \geq 0$,
  - $S(\varphi) \leq 1$,
  - $\neg(S(\varphi) \leq k \wedge S(\varphi) \geq k')$ for all $k < k'$,
  - $S(\varphi) \geq k \vee S(\varphi) \leq k$,
  - if $\vdash \varphi \leftrightarrow \psi$ then $\vdash S(\varphi) \gtreqless c \leftrightarrow S(\psi) \gtreqless c$ (for $\gtreqless \in \{\geq, \leq\}$),
- necessitation for public announcement: if $\vdash \psi$, then $\vdash [!\varphi]\psi$,
- the reduction axioms for public announcement described in Definition 11.

Note that the static axioms for surprise are all concerned with the technical details of this particular formalization of surprise (such as the totality of $\geq$), rather than with any substantial properties of surprise itself. The only substantial axioms for surprise are thus its reduction axioms (items 6-7 of Definition 11), which together constitute a dynamified version of Macedo and Cardoso's original definition (1). I take this to be a clear manifestion of the essentially dynamic nature of surprise in the axiomatization of the logic.

I will finish this subsection by showing that the logic's semantics and axiomatization are in perfect harmony: the axiom system is sound and complete with respect to the semantics.

Theorem 13. *$\mathcal{SURPRISE}$ is (weakly) sound and complete with respect to $\mathcal{C}_S$.*

Proof. As usual, soundness is proved by induction on derivation length. It is easy to check that all axioms of $\mathcal{SURPRISE}$ are semantically valid on $\mathcal{C}_S$, and that all of its rules are $\mathcal{C}_S$-validity-preserving (for the public announcement necessitation rule, recall Lemma 6).

Completeness can also be proved using standard techniques. First of all, because the reduction axioms allow us to rewrite any formula as an equivalent formula without any dynamic operators, it suffices to prove completeness for the static fragment of the logic. This is done using a filtration of a canonical model over a set of formulas Σ which is finite and closed under subformulas. These methods are well-known in probabilistic epistemic logic [15], so I will only discuss the surprise component.

The following can easily be proved for maximally consistent sets $\Gamma \subseteq \Sigma$:

- for all $\chi \in \Sigma$, there exists a number $\alpha_{\Gamma,\chi} \in [0,1] \cap \mathbb{Q}$ such that the formula $S(\chi) = \alpha_{\Gamma,\chi}$ is consistent with Γ,
- for all $\chi, \chi' \in \Sigma$, if $\vdash \chi \leftrightarrow \chi'$, one can always choose $\alpha_{\Gamma,\chi} = \alpha_{\Gamma,\chi'}$.

The canonical model $\mathbb{M}^c$ has states $W^c = \{\Gamma \subseteq \Sigma \mid \Gamma \text{ is maximally consistent}\}$; its surprise function is defined as follows: for all $\Gamma \in W^c$ and $X \subseteq W^c$, put

$$\sigma^c(\Gamma)(X) := \begin{cases} \alpha_{\Gamma,\chi} & \text{if } \exists \chi \in \Sigma : X = \{\Delta \in W^c \mid \chi \in \Delta\}, \\ 0 & \text{otherwise.} \end{cases}$$

The truth lemma can now easily be extended to the case of surprise formulas. Suppose, for example, that the formula $S(\chi) \geq c$ belongs to Σ; then χ itself also belongs to Σ, and by the definition of σ^c, showing that $\mathbb{M}^c, \Gamma \vDash S(\chi) \geq c$ iff $S(\chi) \geq c \in \Gamma$ boils down to showing that $\alpha_{\Gamma,\chi} \geq c$ iff $S(\chi) \geq c \in \Gamma$. The latter follows from the fact that the formula $S(\chi) = \alpha_{\Gamma,\chi}$ is consistent with Γ. □

Corollary 14. *$\mathcal{SURPRISE}$ has the finite model property.*

Proof. Trivial, since surprise models are, by definition, finite. □

4.2. *Some Interesting Results*

I will now show that the logical system developed in the previous subsection is able to capture several properties of surprise. However, there is one technical caveat. Recall that φ can only be publicly announced if φ is true *before* the announcement. It is natural to assume that φ will still be true *after* the announcement. However, because public announcements take into account higher-order information, it might happen that φ, simply by being announced, becomes false. A typical example is $\varphi = p \wedge \neg Kp$. If no such 'self-falsifying' effects occur, φ is called successful:

Definition 15. A formula $\varphi \in \mathcal{L}$ is called successful iff $\vDash [!\varphi]\varphi$.

When modeling 'real-life' scenarios in a single-agent setting, formulas typically do not involve higher-order information,[22] so at least from this modeling perspective, the assumption of successfulness in many of the propositions below is quite harmless.[23] I now turn to the first concrete result.

Proposition 16. *The following formula is satisfiable:*

$$\begin{array}{lllllllll} \varphi & \wedge & \neg K\varphi & \wedge & P(\varphi)=0.2 & \wedge & S(\varphi)=0 \\ & \wedge & \langle!\varphi\rangle(K\varphi & \wedge & P(\varphi)=1 & \wedge & S(\varphi)=0.8) \\ & \wedge & \langle!\varphi\rangle\langle!\varphi\rangle(K\varphi & \wedge & P(\varphi)=1 & \wedge & S(\varphi)=0). \end{array}$$

Proof. Consider $\mathbb{M} := \langle W, R, \sigma, \mu, V\rangle$, with $W = \{w, v\}$, $R = W \times W$, $V(p) = \{w\}$, $\mu(w)(w) = 0.2, \mu(w)(v) = 0.8$ and $\sigma(w)(X)$ and $\sigma(v)(X)$ undefined for all $X \subseteq w$ (all components which have not been mentioned are irrelevant, and can thus be assigned values at random). One can easily check that this is indeed a surprise model, and that the formula stated above (with φ instantiated to p) is indeed true at $\mathbb{M}, w$. Finally, note that $\mathbb{M} \in \mathcal{C}_S^*$, i.e. we have followed the heuristic rule HEUR. □

Proposition 16 shows that the logic is capable of doing what it was designed to do, viz. explicitly representing surprise dynamics. It describes the following scenario. Initially, φ is true, but the agent does not know this. Furthermore, she assigns rather low prior probability to it (and thus does not expect its announcement). However, because she does not yet know that φ is actually true, she experiences no surprise about it whatsoever. Next, the unexpected announcement of φ occurs, and three things happen: (i) the agent comes to know that φ, (ii) she processes this new information by Bayesian conditionalization and thus assigns probability 1 to it, and (iii) she experiences a very high degree of surprise about φ (inversely correlated to the low probability that she initially assigned to it). After another announcement of φ, the agent's knowledge and probabilities are not changed; however, because this second announcement was no longer unexpected (after all, in the meanwhile she has come to know that φ), her surprise about φ drops again to 0. The formula in Proposition 16 captures this scenario in a very natural way, using nested public announcement operators to explicitly represent the successive layers of surprise dynamics.

[22] In a single-agent setting one is typically surprised about 'facts of nature', not about one's *epistemic attitudes about* such facts. In a multi-agent setting, however, it would be natural to have scenarios like "Alice was surprised when finding out that *Bob knows that* φ".

[23] Next to the 'standard' unsuccessful formulas involving knowledge ($p \wedge \neg Kp$, [50]) and probability ($p \wedge P(p) < 1$, [18]), one can also define unsuccessful formulas involving the surprise operator S, e.g. $P(p) > 0 \wedge S(p) \geq 1$. Clearly, these formulas all have the same underlying syntactic structure.

At this point, it should be pointed out that not all scenarios described by satisfiable formulas are equally plausible. In particular, it is easy to check that formulas of the form $S(\varphi) > 0 \wedge P(\varphi) < 1$ are satisfiable, although the scenario described by such formulas sounds highly counterintuitive: the first conjunct says that the agent experiences some surprise about φ, which normally only happens after a public announcement of φ; however, this announcement should also have led the agent to become certain about φ (i.e. to assign probability 1 to it), which contradicts the second conjunct. To rule out such scenarios, one might consider adding an axiom of the form $S(\varphi) > 0 \rightarrow P(\varphi) = 1$ to the $\mathcal{SURPRISE}$ system. However, this ignores the fact that there exist formulas φ to which the agent does *not* assign probability 1 after they have been announced.[24] Furthermore, even if such formulas are excluded—for example, by only considering (Boolean combinations of) propositional atoms—, then it is still the case that $\mathcal{C}_S \not\models S(p) > 0 \rightarrow P(p) = 1$. However, it *does* hold that $\mathcal{C}_S^* \models S(p) > 0 \rightarrow P(p) = 1$. In other words, even though the formula $S(p) > 0 \wedge P(p) < 1$ is satisfiable in a $\mathcal{C}_S$-model, it is *not* satisfiable in a $\mathcal{C}_S^*$-model. Hence, when we are modeling concrete scenarios (and following the heuristic rule Heur), the entire problem does not arise.[25]

We now turn to Proposition 17 below. This says that an occurrence of φ can lead to surprise about φ itself, but also about all of its consequences. For example, it follows from items 1 and 2 that if an agent assigns probability 0.2 to $p \wedge q$, then after the announcement of this conjunction, she is surprised with intensity 0.8 about $p \wedge q$, but also about p and q individually. Items 3 and 4 are trivial consequences of 1 and 2; they are mentioned to highlight the subtleties of unsuccessful formulas: if φ is not assumed to be successful, then 4 continues to hold, but 3 doesn't.

Proposition 17. *Assume that $\varphi \in \mathcal{L}$ is successful, and that $\models \varphi \rightarrow \psi$. Then:*

1. $\models P(\varphi) \geq c \rightarrow [!\varphi]\, S(\psi) \leq 1 - c$,
2. $\models P(\varphi) \leq c \rightarrow [!\varphi]\, S(\psi) \geq 1 - c$,
3. $\models P(\varphi) \geq c \rightarrow [!\varphi]\, S(\varphi) \leq 1 - c$,
4. $\models P(\varphi) \leq c \rightarrow [!\varphi]\, S(\varphi) \geq 1 - c$.

Proof. Straightforward applications of the semantics. □

The fact that an occurrence of φ can lead to surprise about its consequences presupposes that the agent is actually able to *draw* those consequences (if

[24] For example, let $\varphi := p \wedge P(p) = 0.05$, and let $\mathbb{M}$ be the surprise model defined in Example 10; it is now easy to check that $\mathbb{M}, w \models \varphi \wedge [!\varphi]\, P(\varphi) = 0$.

[25] Thanks to an anonymous referee for extensive discussion about this.

the agent did not realize that ψ is a logical consequence of φ, then an unexpected occurrence of φ would cause her to be surprised about φ, but not about ψ). In other words, Proposition 17 shows that the logical system assumes the agent to be logically omniscient.[26] An even clearer illustration of this assumption is provided by item 1 of Proposition 18 below, which says that the agent is never surprised about semantic validities. Similarly, items 2 and 3 say that if an agent already knows φ, or assigns probability 1 to it, then she will not be surprised about it. These principles are clearly false for actual human beings, which are not logically omniscient, and can thus e.g. be genuinely surprised upon learning (that some formula is actually) a semantic validity; rather, the main importance of item 1 is that it elucidates Wittgenstein's famous anti-psychologistic claim that "there can never be surprises in logic" [51, Proposition 6.1251].

Proposition 18. *Assume $\varphi \in \mathcal{L}$ is successful. Then:*

1. *if* $\vDash \varphi$, *then* $\vDash [!\varphi]\, S(\varphi) = 0$,
2. $\vDash P(\varphi) = 1 \rightarrow [!\varphi]\, S(\varphi) = 0$,
3. $\vDash K\varphi \rightarrow [!\varphi]\, S(\varphi) = 0$.

Proof. Straightforward applications of the semantics. □

I will finish this subsection by proving two more substantial results, both of which illustrate how important empirical properties of surprise can be obtained as semantic validities of the logical system.

Proposition 19. *Assume $\varphi \in \mathcal{L}$ is successful. Then for all $n \geq 2$, we have:*[27]

$$\vDash [!\varphi]^n S(\varphi) = 0.$$

Proof. First of all, note that since φ is successful, it holds that $\vDash \varphi \leftrightarrow \langle !\varphi \rangle \varphi$; call this principle (†). Consider an arbitrary surprise model $\mathbb{M} = \langle W, R, \mu, \sigma, V \rangle$ and state w, and assume that $\mathbb{M}, w \vDash \varphi$. For any $n \geq 0$, we abbreviate

$$\langle W^n, R^n, \mu^n, \sigma^n, V^n \rangle = \mathbb{M} \restriction n := (\cdots(\underbrace{\mathbb{M} \restriction \varphi) \restriction \varphi \cdots) \restriction \varphi}_{n \text{ times}}.$$

[26] This also illustrates the thoroughly *epistemic* character of surprise: the problem of logical omniscience originally is a problem for epistemic logic, but it automatically carries over into the surprise component.

[27] $[!\varphi]^n$ is defined inductively: $[!\varphi]^0 \psi := \psi$, and $[!\varphi]^{n+1} \psi := [!\varphi][!\varphi]^n \psi$.

Let's now show that $\mathbb{M}, w \models [!\varphi]^{n+1} P(\varphi) = 1$ for all $n \geq 0$. This follows directly from the following calculation:

$$\begin{aligned}
\mu^{n+1}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n+1}) &= \mu^{n+1}(w)([\![\langle!\varphi\rangle\varphi]\!]^{\mathbb{M}\restriction n}) \\
&= \mu^{n+1}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n}) && (\dagger) \\
&= \mu^{n}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n} \mid [\![\varphi]\!]^{\mathbb{M}\restriction n}) = 1. && \text{(Definition 4)}
\end{aligned}$$

We now use this to justify the (‡)-labeled step in the following calculation:

$$\begin{aligned}
\sigma^{n+2}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n+2}) &= \sigma^{n+2}(w)([\![\langle!\varphi\rangle\varphi]\!]^{\mathbb{M}\restriction n+1}) \\
&= \sigma^{n+2}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n+1}) && (\dagger) \\
&= 1 - \mu^{n+1}(w)([\![\varphi]\!]^{\mathbb{M}\restriction n+1}) && \text{(Definition 4)} \\
&= 1 - 1 = 0. && (\ddagger)
\end{aligned}$$

This shows that $\mathbb{M}, w \models [!\varphi]^{n+2} S(\varphi) = 0$ for all $n \geq 0$. □

Informally speaking, Proposition 19 says that after two public announcements of φ, the agent is no longer surprised about φ. It thus nicely captures the transitory nature of surprise, which was discussed in Subsection 2.1. Furthermore, the proof closely resembles the informal explanation which was given there: the first announcement of φ causes the agent to update her probabilities and to assign probability 1 to φ, so that the second (and subsequent) announcement is no longer unexpected, and thus no longer surprising.[28]

Finally, Proposition 20 says that if an occurrence of (a public announcement of) φ leads an agent to change her probability of ψ from a to b in a non-trivial[29] fashion, then she will experience at least *some* surprise about ψ. In other words: surprise is a *necessary condition* for belief revision (in the

[28] The fact that surprise intensity drops to 0 after only two announcements is no problem for Proposition 19, even though for most real subjects this drop happens more gradually and requires several more repetitions [4]. The more gradual decrease in surprise intensity is the consequence of personal and coincidental factors, such as intelligence and fatigue. Both the informal explanation in Subsection 2.1 and Proposition 19 make abstraction of such factors, and predict that the drop in surprise intensity will already happen after the second repetition.

[29] This non-triviality requirement is captured by the condition that $\models \neg\psi \rightarrow [!\varphi]\neg\psi$, i.e. the public announcement of φ should not turn any $\neg\psi$-states into ψ-states. In other words, the change of $P(\psi)$ from a to b is non-trivial if $[\![\psi]\!]^{\mathbb{M}}$ does not grow. (If $[\![\psi]\!]^{\mathbb{M}}$ grows, then it is trivial that the value of $P(\psi)$ might change: if $A \subseteq B$, then $P(A) \leq P(B)$.) Intuitively, exactly the same argument can be made about $[\![\psi]\!]^{\mathbb{M}}$ shrinking rather than growing (i.e. about the requirement that $\models \psi \rightarrow [!\varphi]\psi$), but it turns out that this second requirement is technically speaking not necessary for Proposition 20 to hold. This disanalogy is similar to the disanalogy between items 3 and 4 of Proposition 17.

current framework: probability revision).[30] This is perfectly in line with the cognitive-psychoevolutionary theory of surprise described in Subsection 2.1, which holds that surprise is part of a sequence of processes triggered by an unexpected event; the final stage of this sequence is typically a process of belief revision.

Proposition 20. *Consider $\varphi, \psi \in \mathcal{L}$ and suppose that $\models \neg\psi \to [!\varphi]\neg\psi$. Then*

$$\models (P(\psi) = a \wedge [!\varphi]P(\psi) = b \wedge a \neq b) \;\to\; [!\varphi]S(\psi) > 0.$$

Proof. Consider an arbitrary surprise model $\mathbb{M} = \langle W, R, \mu, \sigma, V \rangle$ and state w, and assume that the antecedent of the formula above is true at $\mathbb{M}, w$. For a reductio, assume that $\mathbb{M}, w \not\models [!\varphi]\, S(\psi) > 0$. Then it follows that

$$0 = \sigma^{\varphi}(w)([\![\psi]\!]^{\mathbb{M}\restriction\varphi}) = \sigma^{\varphi}(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}}) = 1 - \mu(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}}),$$

and thus $\mu(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}}) = 1$. From the assumption that $\models \neg\psi \to [!\varphi]\neg\psi$ in the statement of the proposition, it follows that $[\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}} \subseteq [\![\psi]\!]^{\mathbb{M}}$, and thus

$$1 = \mu(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}}) \leq \mu(w)([\![\psi]\!]^{\mathbb{M}}) = a,$$

so $a = 1$. Since $\models \langle!\varphi\rangle\psi \to \varphi$, we similarly get that $\mu(w)([\![\varphi]\!]^{\mathbb{M}}) = 1$, and hence

$$b = \mu^{\varphi}(w)([\![\psi]\!]^{\mathbb{M}\restriction\varphi}) = \mu^{\varphi}(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}}) = \frac{\mu(w)([\![\langle!\varphi\rangle\psi]\!]^{\mathbb{M}})}{\mu(w)([\![\varphi]\!]^{\mathbb{M}})} = \frac{1}{1} = 1.$$

We thus have $a = 1 = b$, which contradicts the assumption that $a \neq b$. □

Corollary 21. *For any $\varphi \in \mathcal{L}$, it holds that*

$$\models (P(\varphi) = a \wedge [!\varphi]P(\varphi) = b \wedge a \neq b) \;\to\; [!\varphi]S(\varphi) > 0.$$

Proof. It always holds that $\models \neg\varphi \to [!\varphi]\neg\varphi$, so by putting $\psi = \varphi$, the condition of Proposition 20 is always satisfied. □

5. Conclusion

In this paper I have presented a new analysis of surprise in the framework of probabilistic dynamic epistemic logic. This analysis is based on current

[30] I use the term 'belief revision' in a strictly technical sense here (i.e. as synonymous to 'probability revision'), and do not mean to suggest any connection with AGM-style theories of belief revision [1, 16].

psychological theories, and as a result, several experimentally observed aspects of surprise can be derived as theorems within the logical system (recall, for example, Proposition 20 on the role of surprise in belief revision). Furthermore, being based on the contemporary 'lingua franca' of (dynamic) epistemic logic, it offers a natural, well-understood and highly expressive language for the formal description of agent architectures (cf. Proposition 16).

Most importantly, however, the analysis naturally captures the dynamic nature of surprise. This is clearly manifested in the logic's semantics (the surprise measures $\sigma(w)$ are not required to satisfy any static properties) as well as in its proof theory (the only substantial axioms for surprise are its reduction axioms). These reduction axioms jointly constitute a temporally coherent definition of surprise, in contrast to earlier, temporally incoherent formalizations such as Macedo and Cardoso's and Lorini and Castelfranchi's. This temporal coherence has several advantages. First and foremost, by explicitly distinguishing between prior and posterior notions, the proposed analysis is able to reach a high level of *conceptual hygiene* (recall the methodological remark at the beginning of Subsection 3.2). This conceptual advantage also yields additional *empirical* benefits: the new analysis can capture important aspects of surprise that are not covered by earlier frameworks, such as its transitory nature (cf. Proposition 19).[31]

Several questions are left for further research. For example, I intend to explore what happens with the propositions mentioned in Subsection 4.2 when the assumption of successfulness is lifted (unsuccessful formulas require higher-order information, and thus seem to arise most naturally in multi-agent scenarios; cf. Footnote 4.2). Another topic involves adding awareness to the logic, which would greatly increase its empirical adequacy (cf. Proposition 18). Finally, one might wonder whether the quantitative notion of surprise intensity can be used to define a qualitative notion of surprise, just like in the epistemic realm one can use probabilities (degrees of belief) to define a qualitative notion of belief.[32] These questions, however, will be addressed in another paper.

Lorenz Demey
Institute of Philosophy - CLAW
KULeuven
lorenz.demey@hiw.kuleuven.be

[31] Unsurprisingly, the aspect of transitoriness is itself of a highly dynamic character, involving repeated occurrences of the unexpected event.

[32] In philosophy this move is often called the 'Lockean thesis'; it yields a probabilistic notion of belief which has exactly the same dynamics under public announcement as a 'primitively qualitative' notion of belief [12].

References

[1] Alchourrón, C., Gärdenfors, P., and Makinson, D. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic 50* (1985), 510–530.

[2] Bates, J. The role of emotion in believable agents. *Communications of the ACM 37*, 7 (1994), 122–125.

[3] Becker, C. W., Kopp, S., and Wachsmuth, I. Simulating the emotion dynamics of a multimodal conversational agent. In *Affective Dialogue Systems (ADS 2004)*, E. André, L. Dybkjær, W. Minker, and P. Heisterkamp, Eds., Lecture Notes in Computer Science 3068. Springer, Berlin, 2004, pp. 154–165.

[4] Charlesworth, W. R. Instigation and maintenance of curiosity behavior as a function of surprise versus novel and familiar stimuli. *Child Development 35* (1964), 1169–1186.

[5] Chow, T. Y. The surprise examination or unexpected hanging paradox. *American Mathematical Monthly 105* (1998), 41–51.

[6] Davidson, D. Rational animals. *Dialectica 36* (1982), 317–327.

[7] Demey, L., Kooi, B., and Sack, J. Logic and probability. In *Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. CSLI, Stanford, CA, 2013.

[8] Demey, L., and Kooi, B. Logic and probabilistic update. In *Johan van Benthem on Logic and Information Dynamics*, A. Baltag and S. Smets, Eds., Dordrecht, Springer, 2014, pp. 381–404.

[9] Demey, L., and Sack, J. Epistemic logic and probabilities. In *Handbook of Epistemic Logic*, H. van Ditmarsch, J. Halpern, W. van der Hoek, and B. Kooi, Eds., London College Publications, 2015, pp. 147–202.

[10] Demey, L. Agreeing to disagree in probabilistic dynamic epistemic logic. Master's thesis, ILLC, Universiteit van Amsterdam, Amsterdam, 2010.

[11] Demey, L. Agreeing to disagree in probabilistic dynamic epistemic logic. *Synthese 191* (2014), 409–438.

[12] Demey, L. Contemporary epistemic logic and the Lockean thesis. *Foundations of Science 18* (2013), 599–610.

[13] El-Nasr, M. S. Modelling emotion dynamics in intelligent agents. Master's thesis, Texas A&M University, College Station, TX, 1998.

[14] Faghihi, U., Poirier, P., and Larue, O. Emotional cognitive architectures. In *Affective Computing and Intelligent Interaction (ACII 2011)*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., Lecture Notes in Computer Science 6974. Springer, Berlin, 2011, pp. 487–496.

[15] Fagin, R., and Halpern, J. Reasoning about knowledge and probability. *Journal of the ACM 41* (1994), 340–367.

[16] Gärdenfors, P. *Knowledge in Flux*. MIT Press, Cambridge, MA, 1988.

[17] Harel, D., Kozen, D., and Tiuryn, J. *Dynamic Logic*. MIT Press, Cambridge, MA, 2000.

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

[19] Lorini, E., and Castelfranchi, C. The unexpected aspects of surprise. *Int. J. of Pattern Rec. and AI 20* (2006), 817–833.

[20] Lorini, E., and Castelfranchi, C. The cognitive structure of surprise: Looking for basic principles. *Topoi 26* (2007), 133–149.

[21] Lorini, E. Agents with emotions: A logical perspective. *Association for Logic Programming Newsletter 21*, 2–3 (2008), 1–9.

[22] Macedo, L., Cardoso, A., Reisenzein, R., Lorini, E., and Castelfranchi, C. Artificial surprise. In *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and AI*, J. Vallverdú and D. Casacuberta, Eds. IGI Global, Hershey, PA, 2009, pp. 267–291.

[23] Macedo, L., Cardoso, A., and Reisenzein, R. Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions. In *Proc. of the 26th An. Conf. of the Cog. Sci. Soc.* (Mahwah, NJ, 2004), K. Forbus, D. Gentner, and T. Regier, Eds., Erlbaum, pp. 588–593.

[24] Macedo, L., Cardoso, A., and Reisenzein, R. A surprise-based agent architecture. In *Proc. of the 18th European Meeting on Cybernetics and Systems Research* (Vienna, 2006), R. Trappl, Ed., Austrian Society for Cybernetic Studies, pp. 583–588.

[25] Macedo, L., and Cardoso, A. Creativity and surprise. In *Proc. of the AISB '01 Symposium on Creativity in Arts and Science* (York, 2001), G. Wiggins, Ed., The Society for the Study of Artificial Intelligence and Simulation Behaviour, pp. 84–92.

[26] Macedo, L., and Cardoso, A. Modelling forms of surprise in an artificial agent. In *Proc. of the 23rd An. Conf. of the Cog. Sci. Soc.* (Edinburgh, 2001), J. Moore and K. Stenning, Eds., Erlbaum, pp. 588–593.

[27] Macedo, L., and Cardoso, A. Exploration of unknown environments with motivational agents. In *Proc. of the Third Int. Joint Conf. on Autonomous Agents and MAS* (New York, NY, 2004), N. Jennings and M. Tambe, Eds., IEEE Computer Society, pp. 328–335.

[28] Macedo, L., Reisenzein, R., and Cardoso, A. Surprise and anticipation in learning. In *Encyclopedia of the Sciences of Learning*, N. M. Seel, Ed. Springer, New York, NY, 2012, pp. 3250–3253.

[29] Marsella, S. C., and Gratch, J. EMA: A process model of appraisal dynamics. *Cognitive Systems Research 10* (2009), 70–90.

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

[31] Miller, S. A. Contradiction, surprise, and cognitive change: the effects of disconfirmation of belief on conservers and nonconservers. *Journal of Experimental Child Psychology 15* (1973), 47–62.

[32] Nagel, T. What is it like to be a bat? *Philosophical Review 83* (1974), 435–450.

[33] Ortony, A., and Partridge, D. Surprisingness and expectation failure: What's the difference? In *Proc. of the 10th Int. Joint Conf. on AI* (Los Altos, CA, 1987), J. McDermott, Ed., Morgan Kaufmann, pp. 106–108.

[34] Peirce, C. S. *Collected Papers, vol. 5: Pragmatism and pragmaticism*. Harvard University Press, Cambridge, MA, 1934.

[35] Peirce, C. S. *Collected Papers, vol. 8: Reviews, correspondence, and bibliography*. Harvard University Press, Cambridge, MA, 1958.

[36] Reisenzein, R., Meyer, W.-U., and Schützwohl, A. Reacting to surprising events: A paradigm for emotion research. In *Proc. of the 9th Conf. of the Int. Soc. for Research on Emotions* (Toronto, 1996), N. Frijda, Ed., ISRE, pp. 292–296.

[37] Reisenzein, R., and Meyer, W.-U. Surprise. In *Oxford Companion to the Affective Sciences*, D. Sander and K. R. Scherer, Eds. Oxford University Press, Oxford, 2009, pp. 386–387.

[38] Reisenzein, R. The subjective experience of surprise. In *The message within: The role of subjective experience in social cognition and behavior*, H. Bless and J. P. Forgas, Eds. Psychology Press, Philadelphia, PA, 2000, pp. 262–279.
[39] Rumelhart, D. E. Schemata and the cognitive system. In *Handbook of Social Cognition*, R. S. W. Jr. and T. K. Srull, Eds. Lawrence Erlbaum, Hillsdale, NJ, 1984, pp. 161–188.
[40] Sack, J. Extending probabilistic dynamic epistemic logic. *Synthese 169* (2009), 241–257.
[41] Schank, R. *Explaining Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum, Hillsdale, NJ, 1986.
[42] Shoham, Y., and Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, Cambridge, 2009.
[43] Sokolov, E. N., Spinks, J. A., Näätänen, R., and Lyytinen, H. *The orienting response in information processing*. Lawrence Erlbaum, Mahwah, NJ, 2002.
[44] Stiensmeier-Pelster, J., Martini, A., and Reisenzein, R. The role of surprise in the attribution process. *Cognition and Emotion 9* (1995), 5–31.
[45] Talbott, W. Bayesian epistemology. In *Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. CSLI, Stanford University, Stanford, CA, 2008.
[46] van Benthem, J., Gerbrandy, J., and Kooi, B. P. Dynamic update with probabilities. *Studia Logica 93* (2009), 67–96.
[47] van Benthem, J., Pacuit, E., and Roy, O. Towards a theory of play: A logical perspective on games and interaction. *Games 2* (2011), 52–86.
[48] van Benthem, J. *Exploring Logical Dynamics*. CSLI Publications, Stanford, CA, 1996.
[49] van Benthem, J. *Logical Dynamics of Information and Interaction*. Cambridge University Press, Cambridge, 2011.
[50] van Ditmarsch, H., van der Hoek, W., and Kooi, B. P. *Dynamic Epistemic Logic*. Springer, Dordrecht, 2007.
[51] Wittgenstein, L. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1922.
[52] Wooldridge, M. *An Introduction to Multiagent Systems*. John Wiley & Sons, West Sussex, 2002.