Artificial Surprise

Luis Macedo, Amilcar Cardoso, Rainer Reisenzein, Emiliano Lorini, Cristiano Castelfranchi

Chapter XV
Artificial Surprise

Luis Macedo
University of Coimbra, Portugal

Amilcar Cardoso
University of Coimbra, Portugal

Rainer Reisenzein
University of Greifswald, Germany

Emiliano Lorini
Institute of Cognitive Sciences and Technologies, Italy & Institut de Recherche en Informatique de Toulouse, France

Cristiano Castelfranchi
Institute of Cognitive Sciences and Technologies, Italy

ABSTRACT

This chapter reviews research on computational models of surprise. Part 1 begins with a description of the phenomenon of surprise in humans, reviews research on human surprise, and describes a psychological model of surprise (Meyer, Reisenzein, & Schützwohl, 1997). Part 2 is devoted to computational models of surprise, giving special prominence to the models proposed by Macedo and Cardoso (e.g., Macedo & Cardoso, 2001b) and by Lorini and Castelfranchi (e.g., Lorini & Castelfranchi, 2007). Part 3 compares the two models of artificial surprise with each other and with the Meyer et al. model of human surprise, discusses possible targets of future research, and considers possible practical applications.

INTRODUCTION

Considered by some theorists a biologically basic emotion (e.g., Izard, 1991), surprise has long been of interest to philosophers and psychologists. In contrast, the artificial intelligence and computational modeling communities have until recently largely ignored surprise (for an exception, see Ortony & Partridge, 1987). However, during the last years, several computational models of surprise, including concrete computer implementations, have been developed. The aim of these computational models of surprise—which are in part based on psychological theories and findings on the subject—is on the one hand to simulate surprise in order to advance the understanding of surprise in humans, and on the other hand to provide artificial agents (softbots or robots) with the benefits of a surprise mechanism. This second goal is motivated by the belief that surprise is as relevant for artificial agents as it is for humans. Ortony and Partridge (1987, p. 108), proposed that a surprise mechanism is "a crucial component of general intelligence". Similarly, we propose that a surprise mechanism is an essential component of any anticipatory agent that, like humans, is resource-bounded and operates in an imperfectly known and changing environment. The function of the surprise mechanism in such an agent is the same as in humans: To promote the short- and long-term adaptation to unexpected events (e.g., Meyer et al., 1997). As will be seen, this function of surprise entails a close connection of surprise to curiosity and exploration (Berlyne, 1960), as well as to belief revision and learning (e.g., Charlesworth, 1969). Beyond that, surprise has been implicated as an essential element in creativity, aesthetic experience, and humor (e.g., Boden, 1995; Huron, 2006; Schmidhuber, 2006; Suls, 1971). Surprise is therefore also of importance to artificial intelligence researchers interested in the latter phenomena (Macedo & Cardoso, 2001a, 2002; Ritchie, 1999).

The chapter comprises three sections. Section 1 reviews psychological research on surprise. After a brief historical survey, the theory of surprise proposed by Meyer et al. (1997) is described in some detail. Section 2 is devoted to computational models of surprise, giving special prominence to the models of Macedo and Cardoso (e.g., Macedo & Cardoso, 2001b; Macedo et al., 2004) and Lorini and Castelfranchi (e.g., Lorini & Castelfranchi, 2007). Section 3 compares the two models of artificial surprise with each other and with the Meyer et al. (1997) model of human surprise, discusses possible targets of future research, and considers possible practical applications.

SURPRISE IN HUMANS

Pre-Theoretical Characterization of Surprise

Common-sense psychology conceptualizes surprise as a peculiar state of mind, usually of brief duration, caused by unexpected events of all kinds. *Subjectively* (i.e., from the perspective of the surprised person), surprise manifests itself centrally in a phenomenal experience or "feeling" (Reisenzein, 2000b) with a characteristic quality, that can vary in intensity (e.g., one can feel slightly, moderately or strongly surprised). In addition, the surprised person is often aware, at least if she observes herself carefully, of a variety of surprise-related mental and behavioral events: She realizes that something is different from usual or other than expected; she notices that her ongoing mental processes and actions are being interrupted and that her attention is drawn to the unexpected event; she may feel curiosity about the nature and causes of this event; and she may notice the occurrence of spontaneous epistemic search processes (for empirical evidence see e.g., Reisenzein, Bördgen, Holdtbernd, & & Matz, 2006).

Objectively (i.e., from the perspective of the outside observer), surprise may reveal itself—depending on circumstances—in any of a number of behavioral indicators, including: Interruption or delay of ongoing motor activities; orienting of the sense organs to the surprising event; investigative activities such as visual search and questioning others; spontaneous exclamations ("Oh!") and explicit verbal proclamations of being surprised; and a characteristic facial expression consisting, in full-blown form, of eyebrow-raising, eye-widening, and mouth-opening/jaw drop (Ekman, Friesen, & Hager, 2002). Furthermore, psychophysiological studies suggest that surprising events may elicit a variety of bodily changes, commonly subsumed under the so-called *orienting response* (Sokolov, 1963), such as a temporary slowing of heart rate and an increased activity of the eccrine sweat glands (see Meyer & Niepel, 1994). It must be emphasized, however, that the behavioral manifestations of surprise occur by no means in all situations and are in general only loosely associated with one another (Reisenzein, 2000a).

History of Research on Surprise

Descriptions of surprise as a mental and behavioral phenomenon, as well as first attempts at theory-building, date back as far as Aristotle (about 350 B.C.). Among the first to discuss surprise in modern times were the philosophers Hume (1739/1978) and Smith (1795/1982). Their ideas were taken up and elaborated further when psychology was established as an independent discipline in the second half of the 19th century, by authors such as Darwin, (1872/1965), McDougall (1908/1960), Ribot (1896) Shand (1914), Wundt (1863). It is probably fair to say that by 1920, most of the questions of surprise research that can be asked from a noncomputational perspective had been formulated; in addition, first experimental studies of surprise had been conducted. Thus, in a historical survey of surprise research published in 1939, Desai (1939) lists the following issues as having been topics of reflection (plus some empirical research): The elicitors of surprise; the subjective experience of surprise (its nature, feeling tone, and duration); the inhibitory effect of surprise; surprise and attention; surprise and memory; the expression of surprise; surprise and related mental states (e.g., wonder and curiosity); the question of whether surprise is an emotion; the biological function and phylogenetic development of surprise; the ontogenetic development of surprise; the role of surprise in pathology; and the place of surprise in social psychology.

During the behaviorist era of psychology (about 1920-1960), research on surprise came largely to a standstill, to be taken up again only following the "cognitive revolution" of the 1960s. At that time, aspects of surprise first came to be discussed again under the headings of "orienting reaction" (Sokolov, 1963) and "curiosity and exploration" (Berlyne, 1960). Surprise as an independent phenomenon was first discussed anew by evolutionary emotion theorists (Izard, 1971; Tomkins, 1962). Referring back to Darwin (1872/1965), these authors proposed that surprise is a basic emotion that serves essential biological functions. One of these functions—surprise as an instigator of epistemic (specifically causal) search and a precondition for learning and cognitive development—came to be particularly emphasized by developmental psychologists (see Charlesworth, 1969). In the 1970s and 1980s, this suggestion was taken up by social psychologists interested in everyday causal explanations, who emphasized unexpectedness as a main instigator of causal search (e.g., Pyszczynski & Greenberg, 1987; Weiner, 1985). In the 1980s, cognitive psychologists (e.g., Kahneman & Tversky, 1982; Rumelhart, 1984), including cognitively oriented emotion theorists (e. g., Meyer, 1988; Ortony, Clore, & Collins, 1988) became interested in surprise. Since that time, research on surprise as an independent phenomenon has steadily increased and is carried out today by researchers in different subfields of psychology.

Topics addressed by recent psychological research on surprise are, for example, the relation between surprise intensity and the strength of cognitive schemas (e.g., Schützwohl, 1998), the role of surprise in spontaneous attention capture (e.g., Horstmann, 2002), the effects of surprise on the hindsight bias (e. g., Pezzo, 2003), the spontaneous facial expression of surprise (e.g., Reisenzein et al., 2006), and the role of surprise in advertising (e.g., Derbaix & Vanhamme, 2003).

Psychological Theories of Surprise

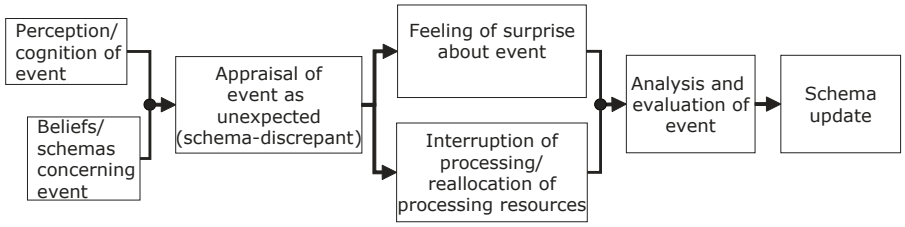
The Cognitive-Psychoevolutionary Model

Classical psychological theories of surprise are formulated exclusively in the language of common-sense psychology, using concepts such as *belief*, *expectation*, *attention* and, of course, *surprise*. In other words, these theories are formulated on what Dennett (1987) called the *intentional level* of system analysis. Only in recent times have there been attempts to move below the intentional level to the *design level* (Dennett, 1987), the level of underlying mental mechanisms, or the cognitive architecture (e.g., Meyer et al., 1997). The aim of these newer "process models" of surprise is to provide a deepened understanding of the causal generation of surprise, its nature, and its functional role in the architecture of the mind, by describing the information-processing mechanisms that underlie the feeling of surprise and surprise-related mental events and behaviors. Although these process models of surprise are not yet detailed enough to count as computational models, they are natural precursors to such models because, although they leave open many issues, they provide enough detail to serve as reasonable starting points for computational modeling. As such, these process models play the role of an intermediary between intentional-level theories of surprise and full-fledged computational theories.

To illustrate recent psychological theorizing surprise, we describe the so-called *cognitive-psychoevolutionary model of surprise* proposed by Meyer et al. (1997); see also, Meyer, Reisenzein, & Niepel (2000). This model is intended as an integration and elaboration of the modal views of previous surprise theorists and attributional analyses of reactions to unexpected events (e.g., Pyszczynski & Greenberg, 1987), within the framework of schema theory (Rumelhart, 1984; Schank, 1986). The model is depicted (in simplified form) in Figure 1.

Schemas as representational structures. Schema theory (e.g., Rumelhart, 1984; Schank, 1986) assumes that human perception, thought and action are to a large extent controlled by complex, organized knowledge (or belief) structures, called schemas. Schemas can be regarded as informal, unarticulated theories, or as sets of beliefs, about objects, events, event sequences (including actions and their consequences) and situations. Schemas serve the interpretation of present and past, and the prediction of future events, and thereby the adaptive guidance of action. To be able to fulfill these functions, a person's schemas (her informal theories) must be at least approximately correct. This in turn requires—because knowledge

Figure 1. The cognitive-evolutionary model of surprise



about the environment is frequently incomplete, and because the environment can change—that schemas are continuously monitored for their compatibility with newly acquired information and, if necessary, are appropriately updated. According to Meyer (1988), Meyer et al. (1997) and Meyer et al. (2000), the surprise mechanism plays a crucial role in this context.

The surprise mechanism is assumed to consist at its core of a device that continuously compares, at an unconscious level of processing, the currently activated cognitive schemas (which may be regarded as constituting the person's working-memory model of her present situation) with newly acquired information (beliefs). As long as this mechanism registers congruence between schema and input—as long as events conform to expectations—the person's informal theories are supported by the evidence, and there is hence no need to revise them. Rather, the interpretation of events and the control of action take place largely automatically and without effort. In contrast, if a discrepancy between schema and input is detected, a "surprise reaction" is elicited (see Figure 1): Ongoing information processing is interrupted, processing resources are reallocated to the unexpected event, surprise is experienced, and cognitive processes (as well as, possibly, overt actions) aimed at the analysis and evaluation of the unexpected event are initiated. The function of these processes is, on the one hand, to enable and motivate immediate adaptive actions directed at the surprising event (*short-term adaptation*); and on the other hand, to promote the appropriate revision of the disconfirmed schemas and thereby, future adaptive actions (*long-term adaptation*).

Surprise processes: A four-step sequence. In more detail, (Meyer et al., 2000; Meyer et al., 1997) assume that (ultimately) surprise-eliciting events elicit a four-step sequence of processes. The first step in this sequence consists of (1) the appraisal of an event as schema-discrepant, or unexpected.[1] If the degree of schema-discrepancy (unexpectedness) exceeds a certain threshold, then (2) ongoing mental processes are interrupted, attention is shifted to the unexpected event, and surprise is experienced. This second step serves to enable and prepare (3) the analysis and evaluation of the unexpected event plus—if this analysis suggests so—(4) immediate reactions to the unexpected event and/or an updating, extension, or revision of the schema or schemas that gave rise to the discrepancy. Ideally, successful schema change (belief update) enables the person to predict and, if possible, to control future occurrences of the schema-discrepant event; to avoid the event if it is negative and uncontrollable; or to ignore the event if it is irrelevant for action.

The surprise mechanism. The first two steps in the posited series of mental processes (Figure 1) are identified with the workings of the surprise mechanism proper. This mechanism is assumed to be a hardwired information processing device whose evolutionary function is to *detect* schema-discrepant events (step 1), and, if they are detected, to *enable* and *prepare* the processes of event analysis and schema revision (steps 3 and 4) by means of the interruption of ongoing processing, the refocusing of attention, and the system-wide communication of the detection of a schema-discrepancy in the form of the feeling of surprise (step 2). In addition, the feeling of surprise is assumed to provide a motivational impetus for the analysis of the surprising event (Meyer et al., 2000). In accordance with the posited hardwiredness of the schema-discrepancy detector, it is assumed that this mechanism operates at a preconscious level of information processing, where it continuously and automatically (specifically without the person's intention) compares activated cognitive schemas with newly acquired information (perceptions, beliefs).

Event analysis. The analysis and evaluation of surprising events (step 3) is assumed to comprise, in typical cases, the following subprocesses: the verification of the schema discrepancy (did one see or hear right; did one draw the correct conclusion from premises?); the analysis of the causes

of the unexpected event (why did it happen?); the evaluation of the unexpected event's significance for well-being (is the event good or bad, is it dangerous or is it a promise?); and the assessment of the event's relevance for ongoing action (can one ignore the event, or does one need to respond to it?). Additional event appraisals, such as an assessment of the moral significance of another person's unexpected action, may occur in some situations. It is assumed that the processes of event analysis can occur in parallel or sequentially, and that in the latter case, they can occur in different sequences. Furthermore, it is assumed that once the appraisals of an unexpected event (e.g., regarding its causes or action relevance) have been computed, they are stored as part of the schema for this event. As a consequence, the analysis of subsequent instances of the same or similar kinds of events can be substantially abbreviated.

The behavioral manifestations of surprise. The cognitive-psychoevolutionary model of surprise assumes that the described mental processes are causally responsible, alone or in combination, for the various behavioral manifestations of surprise (if they occur): the interruption or delay of ongoing motor activities, investigative activities, facial and verbal expressions of surprise, and peripheral physiological reactions. Some of these behavioral manifestations of surprise are presumably functionless side-effects of the mental surprise processes; for example, the interruption of motor activities is a side-effect of the interruption of the mental processes that control it. However, for the greater part, the behavioral manifestations of surprise are probably adaptive processes that, in one way or another, subserve the major function of the surprise mechanism—the short- and long-term adaptation to unexpected events. For example, investigative motor actions are in the service of epistemic analysis; spontaneous and deliberate surprise vocalizations may serve to communicate one's surprise to others, thereby soliciting their help in explaining an unexpected event; and eyebrow-raising may, as Darwin (1872/1965) argued, facilitate the visual exploration of unexpected events in some situations.

The experience of surprise. Meyer et al. (1997) assume that the immediate output of the schema-discrepancy detector is a nonpropositional signal (i.e., a representation characterized by quality and intensity, but without internal structure; see Oatley & Johnson-Laird, 1987; Picard, 1997; Reisenzein, 2000b) whose intensity codes the degree of schema-discrepancy or unexpectedness. Since the subjective experience or "feeling" of surprise also varies in intensity, and does so in close dependence on the degree of schema-discrepancy, it could simply consist of the conscious awareness of the signal produced by the schema-discrepancy detector. However, the feeling of surprise could include additional elements, such as a direct phenomenal awareness of mental interruption (see Reisenzein, 2000b). As mentioned, the feeling of surprise is thought to have an informational and a motivational function: It informs consciousness about the detection of a schema-discrepancy, and it provides an initial impetus for the analysis of the unexpected event. Because the communication that something unexpected happened elicits curiosity (Berlyne, 1960), the motivational effect of the surprise feeling may be based on its informational effect (Reisenzein, 2000b).

The theoretical definition of surprise. On the basis of the described surprise model, it is possible to replace the pre-theoretical characterization of surprise given at the beginning of this section by a more precise, *theoretical (i.e., theory-based) definition* (Reisenzein, 2007). As evident from Figure 1, according to the cognitive-psychoevolutionary model of surprise, the schema-discrepancy signal is the "causal hub in the wheel of surprise": It is the direct or indirect cause of all subsequent mental processes postulated in the model (the feeling of surprise, interruption and attentional shift, event analysis, and schema revision), as well as of the various external manifestations of surprise. Because of its central causal role, the

schema discrepancy signal suggests itself as the best candidate for the scientific referent of surprise. Hence, the theoretical definition of surprise suggested by the cognitive-psychoevolutionary model is as follows: *Surprise is a nonpropositional signal that is the immediate output of the schema-discrepancy detector.* Note that this signal could remain unconscious, for example if it of very low in intensity. Hence, there could be unconscious surprise.

Surprise and Emotions

Common-sense classifies surprise as an emotion. In contrast, in psychology, the question of whether or not surprise is an emotion remains controversial. Whereas some theorist, particularly those with an evolutionary orientation, consider surprise a biologically basic emotion (e.g., Izard, 1991), other authors deny surprise the status of an emotion (e.g., Ortony et al., 1988). The main reasons for not regarding surprise an emotion are: (a) In contrast to paradigmatic emotions such as joy or fear, surprise does not presuppose the appraisal of the eliciting event as positive (desire-congruent) or negative (desire-incongruent); and corresponding to this, (b) the feeling of surprise is per se hedonically neutral, rather than pleasant or unpleasant. However, it is not clear whether these differences between surprise and paradigmatic emotions are sufficient to exclude surprise from the realm of emotions. For one reason, surprise appears to be an essential ingredient of several unquestioned emotional states, such as disappointment, relief, and shock; and the intensity of most emotions is enhanced if their elicitors occur unexpectedly. For another reason, it has been argued that the cognitive mechanism that produces surprise (presumably, a mechanism that compares newly acquired to pre-existing beliefs) and the mechanism that produces hedonic emotions (presumably, a mechanism that compares new beliefs to existing desires) have similar properties and are closely intertwined in their operation (Reisenzein, 2001, 2009): They are both automatic and unconscious mechanisms that "supervise" representations (beliefs and desires, respectively); they seem to operate in parallel on the same inputs (newly acquired beliefs); and their outputs may be integrated at an unconscious level of processing (e.g., the signals of unexpectedness and desire-incongruence may be unconsciously integrated into the emotion of disappointment). If these assumptions are correct, they would constitute good theoretical reasons for (re-) classifying surprise as an emotion (Reisenzein, 2009). In any case, surprise needs to be considered in theories of emotion, even if it is not regarded as an emotion itself.

ARTIFICIAL SURPRISE

Given the important function played by the surprise mechanism in humans, it seems reasonable to allow artificial agents to take advantage of this mechanism. Indeed, we would argue that to the degree that artificial agents are confronted with "adaptive problems" analogous to those that gave rise to the development of the surprise mechanism in humans—the need to react adaptively in imperfectly known and changing environments—they need to be endowed with a surprise mechanism. In recent years, a number of computational models of surprise have been proposed. In this section, we review two of these models of "artificial surprise", developed respectively by Macedo and Cardoso (e.g., Macedo & Cardoso, 2001a; Macedo et al., 2004) and Lorini and Castelfranchi (e.g., Lorini & Castelfranchi, 2007). Both models of artificial surprise were influenced by psychological theories of surprise (e.g., Meyer et al., 1997), and both seek to capture essential aspects of human surprise. These models are therefore more than distant Artificial Intelligence relatives of the surprise mechanism in humans. Rather, they can be considered as attempted simulations of the human surprise mechanism, even though it

needs to be acknowledged that they are in some respects simplifications, and in other respects idealizations.

Other computational approaches to surprise not reviewed here in detail are (Baldi, 2004; Itti & Baldi, 2006; Peters, 1998). Both approaches focus on the role of surprise in vision (the perception of objects, movements, or scenes), and both are mainly concerned with the first step of the surprise process described earlier, the detection of unexpected events and the computation of surprise intensity. For example, central to Baldi and Itti's surprise model is the proposal to compute surprise intensity as the distance (measured by the Kullback-Leibler divergence) between the prior probability distribution over a set of hypotheses and the posterior distribution resulting from the Bayesian updating of the prior distribution on the basis of new information. This proposal could in principle be incorporated into (modified versions of the) more general surprise models discussed here.

The Macedo-Cardoso Model

The artificial agent. The model of surprise developed by Macedo and Cardoso (2001a, 2004), henceforth abbreviated the M&C model, is integrated into an artificial agent whose central function is to explore the environment. In a typical implementation, the agent explores an artificial environment consisting of buildings located at specific positions, that differ in their structural properties (concerning e.g., the shape of the roof, the door, and the windows) and their functions (e.g., home, hotel, church). The agent's design is similar to the BDI (belief-desire-intention) architecture (Bratman, Israel, & Pollack, 1988; Wooldridge, 2002). The actions of the agent consist of moving around in the environment, that is, visiting objects or places. This behavior is driven by several basic motives (desires), whose satisfaction the agent seeks to maximize by his actions, taking into account what it believes to be true of the environment. To date, up to three basic motives have been considered: the desire to reduce hunger, the desire to satisfy curiosity, and—specific to the M&C agent—the desire to experience surprise (see below for more detail). Knowledge about the environment is acquired by means of simulated sensors that provide information about the distance and the visible properties of the objects in the environment within a certain range. Objects outside the range of vision cannot be seen by the agent, and the function of an object (e.g., in the case of buildings: home, hotel, church) becomes known to the agent only when it visits the location of the object. However, the agent forms expectations about unknown aspects of the objects (specifically their function) on the basis of what it sees, and the information stored in memory. The basic processing cycle of the agent—ignoring, for the moment, the computation of surprise—is as follows: (1) The agent samples information from the sensors and computes the current state of the world (e.g., its own position and the position and nature of the objects). (2) Taking as input the current world state, information stored in memory, and probability theory, the agent computes the possible future world states resulting from each action that it currently can perform. (3) From among these actions, the agent selects one that maximizes its subjective expected utility function (e.g., an action that promises to lead to maximal surprise with minimal energy consumption). (4) The agent executes the selected action.

The surprise model. The M&C surprise model integrated into the described agent architecture was mainly inspired by two sources: the psychological surprise model of Meyer et al. (1997, 2000) (see also, Reisenzein, 2000b) described in the preceding section, and an analysis of the cognitive causes of surprise from a cognitive science perspective proposed by Ortony and Partridge (1987). Specifically, the M&C model is a computational implementation—although with some simplifications and changes—of the Meyer et al. surprise model, that draws on Ortony

and Partridge (1987) for the choice of the agent's knowledge structures and the implementation of the appraisal of unexpectedness. In addition, the M&C model contains some unique assumptions, in particular the assumption that one motive of the agent is to maximize surprise.

In agreement with Ortony and Partridge (1987), the knowledge of the M&C agent is both episodic and semantic in nature. In the implementation described above, episodic knowledge consists of information about the location and the properties of specific buildings in the agent's environment plus, in the case of incomplete knowledge (about a building's function), a probability distribution over the possible functions of the building (home, hotel, church, etc.). Semantic knowledge emerges from episodic knowledge through a process of abstraction, in which similar object representations are merged into a prototype. Although the M&C agent can represent the physical structure of objects either propositionally or analogically, we consider here only propositional (sentence-like, predicate-subject) representations. In the M&C agent, the propositional description of an objects consists of a set of attribute-value pairs.

Computation of surprise. The computation of surprise takes place at the beginning of the second step of the agent's processing cycle. At this point, the newly computed world state (represented as a set of input propositions) is compared to the beliefs stored in memory. The agent is surprised if its beliefs conflict with the input propositions, with the intensity of surprise being a function of belief strength (subjective probability). Following Ortony and Partridge (1987), Macedo and Cardoso distinguish between two kinds of beliefs—active and passive expectations—that may be disconfirmed by an input proposition, and accordingly, between two main sources of surprise. For example, in the above-described implementation, whenever the agent perceives a building *g* from a distance (meaning it acquires knowledge of *g*'s structural properties), it computes an *active expectation* concerning the building's function (e.g., "*g* is a hotel with probability .66; *g* is a church with probability .30"). If, upon visiting the building, the agent learns that the building is a post office, it is surprised because its surprise module detects that its active expectations conflict with the new information. This is an example of the first source of surprise, active expectation failure (Ortony & Partridge, 1987). In contrast, when the agent sees a building, it need not have active expectations concerning the building's structural properties. For example, when the agent sees a building with round windows, it need not have computed active expectations concerning the building's windows. Still, by consulting its memory, the agent can infer "after the fact" that, for example, the probability of a rectangular window was .67 and that of a square window was .22. This example illustrates the second source of surprise, surprise due to a conflict of an input proposition with a *passive expectation* (Ortony & Partridge, 1987); that is, a belief that, although not computed prior to the input proposition, was inferred afterwards, as part of the processing of the input.

In the initial version of the M&C model (Macedo & Cardoso, 2001b), the intensity of surprise elicited by an input proposition describing an event *E* (e.g., "building *g* is a hotel") was assumed to be proportional to the degree of unexpectedness of *E*, which was defined as 1-*P*(*E*), the subjective probability of *E*. *P*(*E*), in turn, is computed on the basis of the event frequencies stored in the agent's episodic memory (Macedo & Cardoso, 2001b, 2003). Although there is evidence that supports the assumed, inverse linear relation between subjective probability and surprise (e.g., Reisenzein, 2000a), it can be argued that this function does not correctly predict human surprise in some situations. For example, consider a political election involving three candidates *A*, *B*, and *C* with equal chances of being elected (⅓). Intuitively, one would not feel surprised if *A* (or *B*, or *C*) were elected. To improve this aspect of the model, Macedo et al. (2004) examined several alternative ways of computing surprise intensity

from probability. This study suggested that the relation between subjective probability and the intensity of surprise about an event E_g from a set of mutually exclusive events $\{E_1, E_2, \ldots, E_m\}$ is better described by:

$$SURPRISE(E_g) = \log_2(1 + P(E_h) - P(E_g))$$

In this formula, E_h is the event with the highest probability in the set. The main differences between this surprise intensity function and the simpler function (1-$P(E_g)$) are: (a) Surprise intensity is a *nonlinear* function of probability; (b) the intensity of surprise about an event E_g depends not only on the probability of E_g, but also on that of E_h, the most probable alternative. (The addition of 1 only serves to normalize surprise intensity between 0 and 1). More precisely, the intensity of surprise about E_g is a nonlinear function of the *difference*, or *contrast*, between $P(E_h)$ and $P(E_g)$ (see also, Macedo, Cardoso, & Reisenzein, 2006; Teigen & Keren, 2003). This probability difference can be interpreted as the amount by which the probability of E_g would have to be increased for E_g to become unsurprising. The equation implies that, in each set of mutually exclusive events, there is always at least one event whose occurrence is unsurprising, namely, E_h. As a consequence, one will not be surprised if either one of three equally promising political candidates *A*, *B*, and *C*, is elected.

Computation of expected surprise. The above equation describes the computation of the intensity of *actual surprise* about an event. This computation corresponds to the "appraisal of unexpectedness" in the surprise model of Meyer et al. (1997). However, Macedo and Cardoso assume that the agent in addition computes, for each possible action, the intensity of *expected surprise*, that is, the degree of surprise it will most likely experience if the action is carried out. (In humans, this computation might be performed by means of theoretical inference using a folk theory of surprise; or by means of mental simulation, during which the surprise module is used "off-line"). In the simplest case, the agent's action leads to a future world state *S* in which one of a set of mutually exclusive events $\{E_1, E_2, \ldots, E_m\}$ is realized. For example, if the agent visits a new building, it learns which of the possible functions of buildings this particular building realizes. The degree of surprise expected for *S* is computed analogously to expected utility (e.g., Russell & Norvig, 1995), with surprise intensity taking the place of utility:

$$E[SURPRISE(S)] = \sum_{i=1}^{m} P(E_i) \times \log_2(1 + P(E_h) - P(E_i))$$

Expected surprise resembles the concept of entropy (*H*) in information theory (Shannon, 1948). The difference is that in *H*, surprise intensity as defined here is replaced by "surprisal" (Tribus, 1961), defined as $-\log_2(P(E_i))$.

Computation of the total surprise caused by complex events. So far, we considered only the surprise elicited by a single unexpected event, represented by an input proposition such as "object *g* has a steep roof", or "object *g* is a hotel". However, surprising events are often complex, consisting of several component events. For example, an agent expecting to encounter a hotel with a steep roof may instead find a home with a flat roof. What is the *total surprise* caused by this complex event? The M&C model makes the simplifying assumption that the total surprise elicited by a complex event (e.g., an object with several unexpected features) is the sum of the surprises caused by the different components of the event (Macedo & Cardoso, 2005). Hence, a reductionist approach is taken to the computation of total surprise. To illustrate, assume that the agent is certain that a building *g* at a given distance is a hotel with a steep roof, but then finds out that the building is a home with a flat roof. That is, before encountering object *g*, P(*g is a hotel*) and P(*g has a steep roof*) are both 1, whereas P(*g is a home*) and P(*g has a flat roof*) are both 0. Therefore, *SURPRISE*(*g is a home*) = *SURPRISE*(*g has a flat roof*) = log_2

(1 + 1 − 0) = 1. In the M&C model, the intensity of surprise caused by encountering *a home with a flat roof* is simply the sum of these surprise intensities, i.e., 2. More generally, the intensity of the total (actual) surprise elicited by a complex event *g* consisting of *n* component events E_{1g}, E_{2g}, ..., E_{ng} (here considered as values of dimensions E_1, E_2, ..., E_n) is:

$$SURPRISE(g) = \sum_{j=1}^{n} \log_2 (1 + P(E_{jh}) - P(E_{jg}))$$

Unlike surprise elicited by a single event, the intensity of surprise elicited by a complex event is not normalized (i.e., not limited to 1). This assumption reflects the intuition that the total surprise that can be caused by an object increases with the object's complexity, the number of its different aspects or pieces. That is, other factors constant, more complex objects are potentially more surprising. Still, it might be objected that the formula for total surprise is adequate only if the agent's beliefs about the different components of the complex event are independent (Lorini & Castelfranchi, 2006); whereas, if this is not the case, the formula over- or underestimates total surprise. To illustrate the case of overestimation, assume that the agent in the above example believes not only that a building at a certain distance is a hotel with a steep roof, but also that buildings with flat roofs usually are homes. In this case, upon finding that the building has a flat roof, the agent could immediately revise its belief about the building's function. As a consequence, the agent would be no longer surprised when it learns that the building is a home. Note, however, that this counterexample to the proposed formula for total surprise assumes that the features of the object are processed sequentially. If they are processed in parallel (as seems often plausible to assume for visual perception), then no belief revision can take place. This consideration suggests that the question of the computation of total surprise cannot be fully answered without making assumptions about the parallel versus sequential processing of input propositions (see also, Schimmack & Colcombe, 2007).

Analogous to total actual surprise, the *total expected surprise* for a future situation *X* involving a complex event *g* consisting of *n* component events can be defined as:

$$E\left[SURPRISE(X)\right] =$$

$$\sum_{j=1}^{n} \sum_{i=1}^{m_i} P(E_{ji}) \times \log_2 (1 + P(E_{jh}) - P(E_{ji}))$$

where, for each dimension E_j of the complex event, there are m_i expectations $P(E_{ji})$.

Effects of surprise. As mentioned, the artificial agent into which the M&C surprise model is embedded is driven by several basic motives, including the motive to maximize surprise. The effects of surprise on the agent's actions are easiest to describe if other motives are absent. In this case, following the computation of the actual and anticipated intensity of surprise for each object, the object with the maximum overall (actual plus anticipated) surprise is selected to be visited and investigated. This decision process and the ensuing action simulate aspects of step 2 of the Meyer et al. (1997) model (interruption of ongoing activities and reallocation of processing resources to the surprising event), as well as aspects of step 3 (analysis of the unexpected event). It should be noted, however, that the event analysis in the M&C model is very simple, being restricted to the acquisition of additional information about the object by visiting it. Other aspects of event analysis are currently not considered. Finally, the new information gained about the visited and other objects is stored in episodic memory, and the object frequencies are updated. This is a simplified version of step 4 of the Meyer et al. (1997) model (schema update or belief revision).

The behavior of the M&C surprise agent has been studied in a series of comparative simulations (e.g., Macedo, 2006; Macedo & Cardoso, 2001b, 2004, 2005; Macedo et al., 2004; 2006).

The Lorini-Castelfranchi Model

Theoretical background. The surprise model proposed by Lorini and Castelfranchi (2006, 2007), henceforth abbreviated the L&C model, is part of a more general theory of cognitive expectations and anticipation developed in Castelfranchi (2005) and Miceli & Castelfranchi (2002). In agreement with Macedo and Cardoso (2001b) and Meyer et al. (1997), Lorini and Castelfranchi conceptualize surprise as an expectation- or belief-based cognitive phenomenon, that plays a fundamental role in mental state dynamics. However, different from Macedo and Cardoso and Meyer et al., Lorini and Castelfranchi have explicated their surprise theory as a formal model, using a logic of probabilistically quantified beliefs (Halpern, 2003). An important motive for this formalization was to connect surprise theory to formal models of belief revision in logic and artificial intelligence (e.g., Alchourron, Gärdenfors, & Makinson, 1985; Gerbrandy & Groeneveld, 1997; van Ditmarsch, van der Hoek, & Kooi, 2007). Linking these two research fields seems desirable because, as Lorini and Castelfranchi (2007) point out, formal approaches to belief revision have largely neglected the causal precursors of belief change. However, in contrast to standard models of belief revision, surprise theory (e.g., Meyer et al., 1997) suggests that belief revisions are triggered only under specific conditions, and remain "local" (i.e., only beliefs detected by the surprise mechanism as inconsistent are revised). Hence, surprise theory suggests a strongly "localist" approach to belief revision, that departs from the classical approach (Alchourron et al., 1985) but is close to more recent philosophical work on local belief revision (e.g., Hansson & Wassermann, 2002). Parts of the L&C model of surprise have been implemented in a modified BDI agent (Lorini & Piunti, 2007).

A typology of expectations and forms of surprise. The L&C model distinguishes between several distinct forms of surprise, each of which is based on a different kind of expectation (belief). Specifically, Lorini and Castelfranchi (2006) distinguish *scrutinized expectations* (expectations or beliefs under scrutiny) from *background expectations* (for a similar distinction, see Kahneman & Tversky, 1982). Scrutinized expectations occupy consciousness and draw on the limited capacity of attention. They are anticipatory representations of the next input, which the agent (or a cognitive subsystem) seeks to match to the incoming data, and are closely related to the agent's current intentions and goals. In contrast, *background expectations* reside at an unconscious level of processing. They are either the product of priming (Matt, Leuthold, & Sommer, 1992; Sommer, Leuthold, & Matt, 1998) or part of the background mental framework—the schemas, scripts or knowledge base—that supports the currently scrutinized expectations. The agent's background mental framework includes *conditional expectations*, which constitute the beliefs that the agent uses for interpreting the context in which its action and perception are situated. To illustrate, while trying to find a cheap flight from Rome to London on the Ryanair website, an agent may consciously expect (i.e., may have a scrutinized expectation) to find such a flight there. This scrutinized expectation is supported by a conditional background expectation of the form "If I enter into the Ryanair website, I will find a cheap flight from Rome to London".

Starting from this typology of expectations, Lorini and Castelfranchi (2006, 2007) develop a formal model of surprise that distinguishes between three kinds of surprise: *mismatch-based surprise*, *astonishment*, and *disorientation*.

Mismatch-based surprise. Mismatch-based surprise is surprise caused by a recognized inconsistency between a perceived fact (input proposition) and a scrutinized expectation. In the typical case, the agent has an anticipatory, conscious representation of the next input against which incoming data are matched. Surprise occurs if the agent registers a mismatch between the two

representations. The intensity of mismatch-based surprise depends on the strength of the agent's expectation, defined as the agent's subjective probability of the expected input. More precisely, assume that proposition φ is the (content of a) scrutinized expectation and ψ is the input, and that according to the agent's beliefs, φ and ψ are inconsistent, that is, $\psi \rightarrow \neg\varphi$ (i.e., ψ implies ¬φ). The intensity of mismatch-based surprise caused by the recognition of the inconsistency between the actual input ψ and the expected input φ is then defined as follows:

$$SURPRISE(\psi, \varphi) = k \cdot PROB(\varphi)$$

In this formula, $PROB(\varphi)$ is the agent's subjective probability that φ will occur and k is a weighting factor, i. e. a constant in the interval [0, 1]. Hence, assuming $(\psi \rightarrow \neg\varphi)$, the intensity of mismatch-based surprise about ψ increases linearly with the probability of the expected event φ. The value of k depends on several parameters, including the agent's current motivational dispositions. For instance, k is assumed to be higher when φ is relevant for the agent's goals than when this is not the case (Castelfranchi, 2005).

As an example of mismatch-based surprise, imagine that Mary is waiting for Bob in her office when someone knocks at the door. Mary now forms the scrutinized expectation that φ = *Bob enters the room* (at the next moment). However, at the next moment, when the door opens, Mary sees that ψ = *Bill enters the room*. According to Mary's beliefs, $\psi \rightarrow \neg\varphi$, that is, ψ is inconsistent with Mary's expectation that φ. Registration of this inconsistency causes Mary to feel mismatch-based surprise, whose intensity is proportional to the strength of Mary's belief (i.e., her subjective probability) that φ.

Astonishment. Mismatch-based surprise is surprise caused by an input proposition that is unexpected in the sense of *misexpected*. In contrast, astonishment is surprise caused by an input ψ that is more narrowly speaking *unexpected* in that it does not conflict with a currently scrutinzed expectation of the agent but is inconsistent with the agent's background expectations. The typical case is that of an agent who, while trying to assimilate an input ψ, infers from its background knowledge that the opposite state of affairs ¬ψ is probable and hence, that ψ is improbable. One can also conceive of this case as one where the agent, after the fact, tries to answer the question "Was ψ predictable?" by reconstructing the probability of ψ, and comes to the conclusion that she would rather have expected ¬ψ (see also Ortony & Partridge, 1987; as a limiting case, the agent simply retrieves the previously computed probability of ¬ψ from long-term memory). Assuming the agent believes ¬ψ with subjective probability $PROB(\neg\psi)$, the intensity of astonishment caused by the input proposition ψ is:

$$ASTONISHMENT(\psi) = k \cdot PROB(\neg\psi)$$

where k is again a constant in the interval [0, 1] (cf. the definition of mismatch-based surprise). Thus, the intensity of astonishment about ψ increases linearly with the subjective probability of ¬ψ . Since $PROB(\neg\psi) = 1\text{-}PROB(\psi)$, $ASTONISHMENT(\psi)$ can also be defined as $k \cdot (1\text{-}PROB(\psi))$, that is, as proportional to the degree of improbability of ψ.

Consider again the case where Mary expects Bob to enter her office, but Bill enters instead. As mentioned, in this situation Mary experiences mismatched-based surprise, because her scrutinized expectation that φ = *Bob enters the room* is disconfirmed. In addition, however, Mary may also experience astonishment about ψ = *Bill enters the room*; namely, if ψ conflicts with Mary's background expectations. More precisely, the intensity of Mary's astonishment is proportional to $PROB(\neg\psi)$, where ¬ψ = *Bill does not enter the room*. Note that $PROB(\neg\psi)$ need not be equal to $PROB(\varphi)$, and hence, that the intensity of surprise and astonishment elicited by an input proposition ψ need not be the same. For example, Mary may

consider it fairly probable that Bob will enter her office, but she may be nearly certain that Bill will not enter (since, as she believes, Bill is currently at a congress abroad). As a consequence, Mary will feel more astonished than surprised. In general, *SURPRISE*(ψ, φ) and *ASTONISHMENT*(ψ) will be of equal intensity only if φ and ¬ψ are equivalent for the agent, for only then is *PROB*(φ) = *PROB*(¬ψ) and *PROB*(¬φ) = *PROB*(ψ).

Surprise and astonishment in possibility theory. Alternative definitions of surprise and astonishment become available if one moves beyond the classical, Bayesian analysis of belief strength as subjective probability and enters into the domain of imprecise probabilities and possibility theory (Dubois & Prade, 1988; Shafer, 1976). Here, we consider only the definitions of surprise and astonishment within possibility theory. Although not part of the L&C model of surprise as described in Lorini and Castelfranchi (2007), these definitions are mentioned here because there is some evidence that humans, at least in some situations, reason about uncertainty in accord with possibility theory rather than probability theory (e.g., Raufaste, Da Silva Neves, & Mariné, 2003). In possibility theory, the concept of probability is replaced by the dual concepts of *degree of possibility* and *degree of necessity*. Intuitively, the possibility of a proposition ψ, *POSS*(ψ), is the degree to which ψ is consistent with the agent's background knowledge, whereas the degree of necessity of ψ, *NEC*(ψ), is the degree to which ψ is implied by the agent's background knowledge. Two fundamental assumptions of possibility theory are *NEC*(ψ) = 1-*POSS*(¬ψ), and *NEC*(ψ) ≤ *POSS*(ψ). Moreover, different from Bayesian probability, *NEC*(ψ) + *NEC*(¬ψ) can be < 1. Possibility theory also allows to express the idea of *degree of ignorance* about whether or not ψ is the case. Degree of ignorance is defined as *IGN*(ψ) = 1-(*NEC*(ψ)+*NEC*(¬ψ)). Intuitively, an agent's ignorance about ψ reflects the extent to which the agent's (background) knowledge does not provide sufficient information to allow the agent to infer the exact probability of ψ (see Fig. 2).

Within the framework of possibility theory, the intensity of astonishment caused by the input proposition ψ can be defined as:

$$ASTONISHMENT(\psi) = k \cdot NEC(\neg\psi)$$

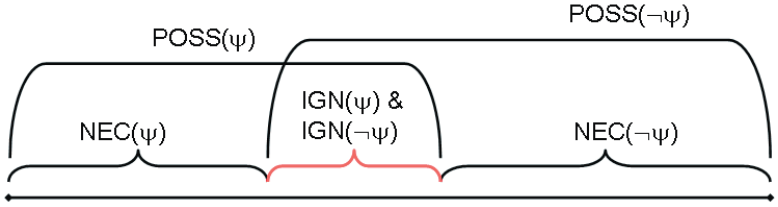
That is, the intensity of astonishment caused ψ is proportional to the degree to which the agent can infer the opposite proposition ¬ψ from its background knowledge. This explication of astonishment corresponds to the concept of *potential surprise* proposed by Shackle (1969). Since *NEC*(ψ) = 1-*POSS*(¬ψ), *ASTONISHMENT*(ψ) can also be defined as $k \cdot (1\text{-}POSS(\psi))$, i. e., as being proportional to the degree of impossibility of ψ.

The intensity of mismatch-based surprise can be defined in possibility theory as:

$$SURPRISE(\psi, \varphi) = k \cdot NEC(\varphi)$$

That is, the intensity of surprise about ψ that conflicts with a scrutinized expectation φ is proportional to the degree to which φ is supported by the agent's background knowledge.

Figure 2. Relation between degree of necessity, possibility, and ignorance



Disorientation. The third form of surprise distinguished in the L&C model is called *disorientation*. Disorientation is surprise caused by the disconfirmation of one or more of the *conditional expectations* that are part of the agent's background knowledge used to interpret the context within which its perception and action are situated. For example, imagine that an agent holding the conditional expectation "If I enter into the Ryanair website, I will find a cheap flight from Rome to London", only finds a flight for 500 Euros on the website. This agent will not only be surprised (because the scrutinized expectation "I will find a cheap flight" is disconfirmed) but probably also disoriented, because the background belief is challenged as well. The agent will then have to reconsider, and possibly to revise this conditional expectation. The intensity of disorientation caused by an input $\neg\psi$ that challenges the conditional expectation "φ entails ψ" is assumed to be proportional to the strength of the expectation, defined by the conditional probability $PROB(\psi \mid \varphi)$.

Functional effects of surprise. The assumptions of the L&C model reviewed so far concern the cognitive origins of surprise. The remaining assumptions of the model concern the functions of surprise in the cognitive system. Similar to the M&C and the Meyer et al. model, it is assumed that surprise serves to suspend the current activity of the agent, to initiate resource mobilization and attention redirection, and to signal a crisis in the assimilation process and the need for accommodation. Particular attention is paid in the L&C model to the role of surprise in the triggering of epistemic processes, including curiosity and exploration, and the instigation of belief revision. Other functions of surprise are suggested in Castelfranchi et al. (2006), where it is proposed, for example, that surprise causes an agent's cautiousness to increase in risky environments.

A surprise-enhanced BDI agent. Parts of the assumptions contained in the L&C model of surprise have been implemented, and in this process further specified, in a computational architecture (Castelfranchi et al., 2006; Lorini & Piunti, 2007). Specifically, the aim was to implement the process of belief change based on mismatch-based surprise. To this end, the authors modified the control loop of a standard BDI agent (Wooldridge, 2002) realized in the JADEX system (Pokahr, Braubach, & Lamersdorf, 2005), by supplementing the JADEX reasoning engine with a filter mechanism for belief revision based on mismatch-based surprise. The resulting, modified BDI architecture embodies two general assumptions. (1) At each moment, the agent focuses attention on a particular task or problem that it is trying to solve. This means that the agent has, at each moment, a set of scrutinized expectations linked to its current intentions and goals. As a consequence, the agent ignores all inputs that are not relevant for the task on which it is currently focused. (2) If a task-relevant input conflicts with a scrutinized expectation of the agent, mismatch-based surprise is elicited, which in turn instigates a belief-update process. In effect, then, mismatch-based surprise signals to the agent that things are not going as expected and that beliefs must be reconsidered.

COMPARISON OF SURPRISE MODELS

As already mentioned in the preceding section, the M&C and L&C models of artificial surprise share quite a few assumptions with each other and with the psychological model of surprise proposed by Meyer et al. However, there are also a number of instructive differences between the models. In this section, the more important similarities and differences are discussed.

1. Both the M&C and the L&C model take as their starting point human surprise, which they seek to model (if in simplified and idealized form) in an artificial agent. Both refer to psychological theories of surprise as a source of inspiration, in particular to

Meyer et al. (1997)), although no attempt is made to include all or only the assumptions of the psychological surprise model.

2. In agreement with most theories of human surprise, both models of artificial surprise conceptualize surprise as a fundamentally expectation- or belief-based cognitive phenomenon, that is, as a reaction to the disconfirmation of expectations or more generally, beliefs. Furthermore, in both models, beliefs are understood as propositional attitudes (e.g., Searle, 1983), and a quantitative belief concept (subjective probability) is used (as an alternative, Lorini and Castelfranchi consider possibility theory).
3. Because beliefs are mental states with propositional contents (Searle, 1983), it is natural to endow the surprise agent with a propositional (sentence-like) medium of representation. In accord with this suggestion, Meyer et al. (1997) propose a schema system (Schank, 1986) to represent belief contents, Macedo and Cardoso (2001b) use a frame-like attribute-value representation, and Lorini and Piunti (2007) take advantage of the frame-like, object-oriented representation of beliefs offered by JADEX. The M&C model also allows for simple analogical representations, but these are not indispensable, nor do they seem to be generally suited to represent the fine-grained, compositional contents of beliefs. Some surprise theorists (e.g., Shand, 1914) have claimed that surprise can also be elicited at "lower" levels of representation than the propositional level, specifically by perceptual mismatch, a possibility also endorsed by Itti and Baldi (2006) and Lorini and Castelfranchi (2007). Presumably, the perceptual applications of their surprise model discussed by Itti and Baldi (2006) are concerned with such "low-level surprise". However, it is doubtful whether perceptual mismatch *per se* causes the experience of surprise in humans (see Niepel, 2001).
4. Both artificial surprise models draw a distinction between two main kinds of expectations or beliefs whose disconfirmation causes surprise (see also, Ortony & Partridge, 1987): Active versus passive expectations (M&C), and scrutinized versus background expectations (L&C). This distinction, whose intent appears to be largely the same in the two models, can be regarded as an attempt to explicate, in computational terms, a distinction frequently drawn in the psychological literature (e.g., Charlesworth, 1969; Meyer, 1988) between two main kinds of unexpected and hence surprise-eliciting events: (a) Events that are *misexpected*, that is, opposite to a prior, specific expectation or belief of the person; and (b) events that are *unexpected in the strict sense*, that is, for which no specific expectation or belief had been inferred at the time when they were cognized (or at least, no such belief was active, i.e. in the agent's working memory), although they conflict with the person's background beliefs. It may be noted in this context that a main reason why Meyer (1988) proposed schemas as the representational structures that underpin surprise was his belief that schema theory (Rumelhart, 1984; Schank, 1986) allows a uniform treatment of both kinds of unexpectedness.
5. In addition to active and passive beliefs, Lorini and Castelfranchi propose conditional expectations (e. g., "If I am on the Ryanair website, I will find a cheap flight") as a third kind of beliefs whose disconfirmation causes surprise. This source of surprise is not explicitly considered as such in the M&C and the Meyer et al. models, although conditional expectations are present in both. In the M&C model, conditional expectations are computed in the process of expectancy generation (Macedo & Cardoso, 2003); in the Meyer et al. model, they are explicitly represented as component of schemas (see also, Lorini &

Castelfranchi, 2007). Furthermore, Meyer et al. emphasize that surprise-caused schema revision concerns frequently not only the revision of the immediately disconfirmed expectation, but also the revision of the more general schemas or "mini-theories" that gave rise to the concrete, disconfirmed expectation. It is plausible to assume that the revision of these more general beliefs, too, is preceded by surprise.

6. Based on the distinction between active (scrutinized) and passive (background) expectations, Lorini and Castelfranchi distinguish between two forms of surprise: Mismatch-based surprise and astonishment. This distinction is not made by Macedo and Cardoso, who speak of "surprise" in both cases. However, inasmuch as the distinction between surprise and astonishment is based on the *causes* of surprise (the disconfirmation of active versus passive expectations), it is implicit in the M&C model. Yet, two differences remain. First, Lorini and Castelfranchi (2007) suggest that their analysis of surprise and astonishment (as well as disorientation) provides for a computational explication of the mental states denoted by these terms in ordinary language. This proposal is ultimately an empirical claim about the referents of the ordinary language terms "surprise", "astonishment" and "disorientation", that could be tested (e.g., Reisenzein, 1995). Second, in contrast to the L&C model, no difference is made in the M&C model between active and passive expectations in the computation of surprise intensity. That is, the intensity of surprise elicited by both kinds of expectation failure is computed according to the same formula, and both contribute equally to total surprise.
7. Although Macedo and Cardoso (2001b) initially used the same surprise intensity function as L&C, according to which the intensity of surprise about an event is proportional to its unexpectedness, Macedo et al. (2004) subsequently opted for a "contrast model" of surprise intensity. This model assumes that the intensity of surprise about an event reflects its probability difference to the contextually most expected event (see also, Reisenzein & Junge, 2006; Teigen & Keren, 2003). Still other probability-based surprise intensity functions have been proposed by other authors (e.g., Itti & Baldi, 2006). Furthermore, Lorini and Castelfranchi suggest an alternative definition of surprise intensity that is based on possibility theory. From the cognitive modeling perspective, the most adequate surprise intensity function is of course the one that best matches that of humans and hence, this difference between the surprise models ultimately needs to be empirically decided. Whether "nonhuman" surprise intensity functions are more useful for *artificial agents* needs to be investigated.
8. Currently, only the M&C model deals explicitly with the question of how to compute the total surprise elicited by a complex event, proposing that is the sum of the surprises elicited by the event's components. However, the L&C model does not preclude the possibility that an event simultaneously disconfirms several expectations, which makes the computation of total surprise a relevant issue. Macedo and Cardoso's proposal concerning the computation of total surprise intensity, while not unproblematic, is the simplest one can make. As mentioned, more sophisticated proposals will likely require to make assumptions about the sequential versus parallel processing of input propositions.
9. The L&C model deals only with the computation of actual surprise, whereas the M&C model also considers the computation of anticipated surprise. That is, M&C propose that the agent uses its surprise module in

two different ways: First, to compute the intensity of actual surprise in response to input; and second, to estimate the intensity of surprise that it will likely "feel" in relevant future situations (those that may result from its actions). Furthermore, only the M&C model assumes that expected surprise influences the agent's cognitions and actions in addition to actual surprise. However, if desired, the L&C model could easily be expanded to include anticipated surprise. It may be noted that the "simulational" use of the surprise module can be extended to predict or explain the surprise of other agents (see Macedo & Cardoso, 2002; Pynadath & Marsella, 2005).

10. Both models of artificial surprise make highly similar assumptions about the functions of surprise in the cognitive architecture, which are in line with the functions of surprise proposed by Meyer et al. (1997), namely: (a) interruption of ongoing activities and focusing of attention to the unexpected event; (b) system-wide communication of the belief- discrepancy and (c) instigation of exploratory activity and belief revision. Castelfranchi et al. (2006) in addition suggest that surprise increases the agent's cautiousness in risky environments.
11. Finally, whereas parts of the L&C model have been implemented in an existing BDI agent architecture, by means of modifying the JADEX reasoning engine, the M&C agent and its surprise module were programmed "from scratch". However, the design of the M&C agent is broadly compatible with the BDI architecture.

FUTURE DEVELOPMENTS

The concrete artificial agents into which the M&C and L&C surprise models have been embedded so far are fairly simple. Although this facilitates the study of the surprise mechanism and its effects on the agents' behavior, future research should also study surprise in agents with more extensive world knowledge and enhanced reasoning capabilities. Only such agents will allow to realistically simulate some surprise-related phenomena in humans, such as surprise-caused causal search (Meyer et al., 1997), the use of the surprise feeling as a source of information in metacognitive reasoning (Reisenzein, 2000b), or the explanation and prediction of surprise in other agents (Pynadath & Marsella, 2005). In addition, to explore the social effects of surprise (e.g., Derbaix & Vanhamme, 2003), future research should study groups of interacting, "surprise-enhanced" agents.

As to the surprise mechanism itself, the comparison of the M&C and L&C models suggests several targets of future research, particularly from the cognitive modeling perspective (i.e., when the models are regarded as simulations of human surprise). For example, are the comparisons of input propositions to active versus passive expectations computed by distinct mechanisms (algorithms), as the L&C model seems to suggest, or by a single mechanism, as the M&C model assumes? Relatedly, if an input conflicts with both an active and a passive expectation, producing both surprise and astonishment (Lorini & Castelfranchi, 2007), does surprise occur first and astonishment later, or do both occur simultaneously? Under which conditions does an agent "reconstruct" the probability of an event that it did not expect? There is also a need for further comparative studies of different surprise intensity functions, as well as for a closer investigation of surprise elicited by complex events (i.e., events with several unexpected aspects). Finally, both artificial surprise models currently lack explicit assumptions about the temporal course of surprise. In particular, does the feeling of surprise spontaneously diminish in intensity according to an intrinsic decay function (e.g., Neal Reilly, 1996), or is surprise reduced only if the responsible schema-discrepancy is resolved or attention shifts elsewhere (also see Pezzo, 2003)?

Going a step further, it would be interesting to expand the artificial surprise models to include other emotions. A straightforward way how this could be achieved is suggested by (Reisenzein, 2009), who sketches a computational model of the belief-desire theory of emotion, a variant of cognitive emotion theory. Following this suggestion, valenced emotions could be incorporated into the surprise models by complementing the mechanism that compares newly acquired beliefs to preexisting *beliefs* (the belief-belief comparator—essentially the surprise mechanism) with another mechanism that compares newly acquired beliefs to preexisting *desires* (the belief-desire comparator). Depending on whether the latter mechanism detects congruence or incongruence between the content of a new belief and that of an existing desires, it produces a feeling of pleasure or displeasure. For details, see Reisenzein (2009).

APPLICATIONS OF ARTIFICIAL SURPRISE

The research on artificial surprise reported in this chapter should be seen in the context of the broader field of affective computing that developed during the past 15 years (Picard, 1997). The aim of affective computing is the computational modeling of emotions, including the expression of emotions and their recognition in other agents. A central motive behind affective computing is the assumption that artificial agents endowed with emotional mechanisms will behave more intelligently than those without. At least in the case of surprise, this assumption is easy to defend. As mentioned, we believe that a surprise mechanism is needed by any resource-bounded anticipatory agent operating in an imperfectly known and changing environment (see also, Ortony & Partridge, 1987).

A second main goal of affective computing is the design of anthropomorphic artificial agents who appear "believable" to human interactants (Bates, 1994), and who adapt their behavior to the interactants' emotions, needs and preferences. Such emotional-expressive agents have manifold possible uses, for example as personal assistants and Embodied Conversational Agents (Cassell, Sullivan, Prevost, & Churchill, 2000), as virtual agents for entertainment (e.g., in games), or as emphatic health care robots. These cognitive agents could profit from artificial surprise research in two ways: First, they could be enhanced by endowing them with a surprise module that influences their actions and belief revision processes. This should not only make the agents more intelligent (Ortony & Partridge, 1987), but also more human-like, by providing them with an emotional state that they can express to humans. Second, because surprise plays an important role in social interaction, artificial agents—even if not "surprise-enhanced" themselves—need a model of human surprise to recognize surprise in their human interaction partners, and to react appropriately to their surprise (e.g., by giving information). Empirical research supports the assumption that intelligent agents who are able to display emotions and to provide emotional feedback to human interactants enhance the users' enjoyment (Prendinger & Ishizuka, 2005) and their evaluation of the artificial agent (Brave, Nass, & Hutchinson, 2005), as well as their engagement (Klein, Moon, & Picard, 1999) and task performance (Partala & Surakka, 2004).

REFERENCES

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

Baldi, P. (2004). Surprise: a shortcut for attention? In L. Itti, G. Rees & J. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 24-28). San Diego, CA: Elsevier Science.

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

Berlyne, D. (1960). *Conflict, arousal and curiosity*. New York: McGraw-Hill.

Boden, M. (1995). Creativity and unpredictability. *Stanford Humanities Review, 4*(2), 123-139.

Bratman, M., Israel, D., & Pollack, M. (1988). Plans and resource-bounded practical reasoning. *Computational Intelligence, 4*(4), 349–355.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*, 161-178.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Castelfranchi, C. (2005). Mind as an anticipatory device: for a theory of expectations. In *Lecture Notes in Computer Science, 3704*, 258-276.

Castelfranchi, C., Falcone, R., & Piunti, M. (2006). Agents with anticipatory behaviors: To be cautious in a risky environment. In *Proceedings of the 17th European Conference on Artificial Intelligence* (*ECAI'06*) (pp. 693-694).

Charlesworth, W. R. (1969). The role of surprise in cognitive development. In D. Elkind & J. H. Flavell (Eds.), *Studies in cognitive development* (pp. 257-314). Oxford: Oxford University Press.

Darwin, C. (1872/1965). *The expression of the emotions in man and animals*. Chicago, IL: University of Chicago Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Derbaix, C., & Vanhamme, J. (2003). Inducing word-of-mouth by eliciting surprise - a pilot investigation. *Journal of Economic Psychology, 24*, 99-116.

Desai, M. M. (1939). Surprise: A historical and experimental study. *British Journal of Psychology, Monograph Supplements, 22*.

Dubois, D., & Prade, H. (1988). *Possibility theory*. New York: Plenum Press.

Ekman, P., Friesen, W. V., & Hager, J. V. (2002). *Facial action coding system* (*2nd Ed.*). Salt Lake City, Utah: Research Nexus eBook.

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

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

Hansson, S. O., & Wassermann, R. (2002). Local change. *Studia Logica, 70*, 49-76.

Horstmann, G. (2002). Evidence for attentional capture by a surprising color singleton in visual search. *Psychological Science, 13*, 499-505.

Hume, D. (1739/1978). *A treatise of human nature*. (*Edited by L. A. Selby-Bigge*). Oxford: Oxford University Press.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge: MIT Press.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems (NIPS 2005), 19*, 1-8.

Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century Crofts.

Izard, C. E. (1991). *The psychology of emotions*. NY: Plenum Press.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*, 143-157.

Klein, J., Moon, Y., & Picard, R. (1999). This computer responds to user frustration. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 242-243). Pittsburgh: ACM Press.

Lorini, E., & Castelfranchi, C. (2006). The unexpected aspects of surprise. *International Journal of Pattern Recognition and Artificial Intelligence, 20*, 817-835.

Lorini, E., & Castelfranchi, C. (2007). The cognitive structure of surprise: looking for basic principles. *Topoi: An International Review of Philosophy, 26*(1), 133-149.

Lorini, E., & Piunti, M. (2007). The benefits of surprise in dynamic environments: from theory to practice. In A. Paiva, R. Prada & R. W. Picard (Eds.), *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction* (Vol. 4738, pp. 362-373). Berlin: Springer.

Macedo, L. (2006). *The exploration of unknown environments by affective agents.* Unpublished PhD, University of Coimbra, Coimbra.

Macedo, L., & Cardoso, A. (2001a). Creativity and surprise. In G. Wiggins (Ed.), *Proceedings of the AISB'01 Symposium on Creativity in Arts and Science* (pp. 84-92). York, UK: The Society for the Study of Artificial Intelligence and Simulation Behaviour.

Macedo, L., & Cardoso, A. (2001b). Modelling forms of surprise in an artificial agent. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 588-593). Edinburgh, Scotland, UK: Erlbaum.

Macedo, L., & Cardoso, A. (2002). Assessing creativity: the importance of unexpected novelty. In *Proceedings of the ECAI'02 Workshop on Creative Systems: Aproaches to Creativity in AI and Cognitive Science,* (pp. 31-37). Lyon, France: University Claude Bernard - Lyon.

Macedo, L., & Cardoso, A. (2003). A model for generating expectations: the bridge between memory and surprise. In C. Bento, A. Cardoso & J. Gero (Eds.), *Proceedings of the 3rd Workshop on Creative Systems: Approaches to Creativity in AI and Cognitive Science, International Joint Conference on Artificial Intelligence* (pp. 3-11). Acapulco, Mexico: IJCAI03.

Macedo, L., & Cardoso, A. (2004). Exploration of unknown environments with motivational agents. In N. Jennings & M. Tambe (Eds.), *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 328 - 335). New York: IEEE Computer Society.

Macedo, L., & Cardoso, A. (2005). The role of surprise, curiosity and hunger on the exploration of unknown environments. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence*. Covilhã, Portugal.

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

Macedo, L., Reisenzein, R., & Cardoso, A. (2004). Modeling forms of surprise in artificial agents: empirical and theoretical study of surprise functions. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 588-593). Mahwah, NJ: Erlbaum.

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

McDougall, W. (1908/1960). *An introduction to social psychology.* London: Methuen.

Meyer, W.-U. (1988). Die Rolle von Überraschung im Attributionsprozes [The role of surprise in the attribution process]. *Psychologische Rundschau, 39*, 136-147.

Meyer, W.-U., & Niepel, M. (1994). Surprise. In V. S. Rachmandran (Ed.), *Encyclopedia of human behavior* (pp. 353-358). Orlando, FL: Academic Press.

Meyer, W.-U., Reisenzein, R., & Niepel, M. (2000). Überraschung [Surprise]. In J. H. Otto, H. A. Euler, & H. Mandl (Eds.), *Emotionspsychologie: Ein Handbuch* (pp. 253-263). Weinheim: Psychologie Verlags Union.

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

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

Neal Reilly, W. S. (1996). *Believable social and emotional agents*. Unpublished PhD Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

Niepel, M. (2001). Independent manipulation of stimulus change and unexpectedness dissociates indices of the orienting response. *Psychophysiology, 38*, 84-91.

Oatley, K., & Johnson-Laird, P. (1987). Towards a cognitive theory of emotions. *Cognition and Emotion, 1*(1), 29-50.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

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

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers, 16*, 295-309.

Peters, M. (1998). Towards artificial forms of intelligence, creativity, and surprise. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (pp. 836-841). Madison, Wisconsin, USA: Erlbaum.

Pezzo, M. V. (2003). Surprise, defence, or making sense: What removes the hindsight bias? *Memory, 11*, 421-441.

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Pokahr, A., Braubach, L., & Lamersdorf, W. (2005). JADEX: a BDI reasoning engine. In R. H. Bordini, M. Dastani, J. Dix & A. El Fallah-Seghrouchni (Eds.), *Multi-agent programming: Languages, platforms and applications* (pp. 149-174). New York: Springer.

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *International Journal of Applied Artificial Intelligence, 19*, 297-285.

Pynadath, D. V., & Marsella, S. (2005). PsychSim: modeling theory of mind with decision-theoretic agents. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 1181-1186).

Pyszczynski, T. A., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model. *Advances in Experimental Social Psychology*.

Raufaste, E., Da Silva Neves, R., & Mariné, C. (2003). Testing the descriptive validity of possibility theory in human judgments of uncertainty. *Artificial Intelligence, 148,* 197-218.

Reisenzein, R. (1995). On Oatley and Johnson-Laird's theory of emotions and hierarchical structures in the affective lexicon. *Cognition and Emotion, 9*, 383-416.

Reisenzein, R. (2000a). Exploring the strength of association between the components of emotion syndromes: The case of surprise. *Cognition and Emotion, 14*, 1-38.

Reisenzein, R. (2000b). The subjective experience of surprise. In H. Bless & J. P. Forgas (Eds.), *The message within: The role of subjective experience in social cognition and behavior* (pp. 262-279). Philadelphia, PA: Psychology Press.

Reisenzein, R. (2001). Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions. In K. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 187-201). Oxford: Oxford University Press.

Reisenzein, R. (2007). What is a definition of emotion? And are emotions mental-behavioral processes? *Social Science Information, 46*, 424-428.

Reisenzein, R. (2009). Emotions as metarepresentational states of mind: Naturalizing the belief-desire theory of emotion. *Cognitive Systems Research, 10*, 6-20.

Reisenzein, R., & Junge, M. (2006). *Überraschung, Enttäuschung und Erleichterung: Emotionsintensität als Funktion von subjektiver Wahrscheinlichkeit und Erwünschtheit* [*Surprise, disappointment and relief: Emotion intensity as a function of subjective probability and desire strength*]. Paper presented at the 45th Congress of the German Psychological Association (DGPs).

Reisenzein, R., Bördgen, S., Holdtbernd, T., & Matz, D. (2006). Evidence for strong dissociation between emotion and facial displays: The case of surprise. *Journal of Personality and Social Psychology, 91*, 295-315.

Ribot, T. A. (1896). *La psychologie des sentiments [The psychology of emotions]*. Paris: Alcan.

Ritchie, G. (1999). Developing the incongruity-resolution theory. In *Proceedings of the AISB. Symposium on Creative Language* (pp. 78-85). Edinburgh, Scotland.

Ruffman, T., & Keenan, T. R. (1996). The belief-based emotion of surprise: The case for a lag in understanding relative to false belief. *Developmental Psychology, 32*, 40-49.

Rumelhart, D. E. (1984). Schemata and the cognitive system. In R. S. Wyer Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (pp. 161-188). Hillsdale, NJ: Erlbaum.

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Schank, R. (1986). *Explanation patterns: understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Schimmack, U., & Colcombe, S. (2007). Eliciting mixed feelings with the paired-picture paradigm: A tribute to Kellogg (1915). *Cognition and Emotion, 21*, 1546-1553.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science, 18*, 173-187.

Schützwohl, A. (1998). Surprise and schema strength. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1182-1199.

Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

Shackle, G. (1969). *Decision, order and time in human affairs* (2nd ed.). Cambridge, UK: Cambridge University Press.

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

Shand, A. F. (1914). *The foundations of character*. London: Macmillan.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423 and 623-656.

Smith, A. (1795/1982). The history of astronomy. *Essays on philosophical subjects, ed. W. P. D. Wightman & J. C. Bryce, vol. 3 of the Glasgow Edition of the Works and Correspondence of Adam Smith* (pp. 5-32). Indianapolis: Liberty Fund.

Sokolov, E. N. (1963). Higher nervous functions. The orienting reflex. *Annual Review of Physiology, 26*, 545-580.

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

Suls, J. M. (1971). A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The psychology of humor* (pp. 81-100). New York: Academic Press.

Teigen, K. H., & Keren, G. B. (2003). Surprises: Low probabilities or high contrasts? *Cognition, 87*, 55-71.

Tomkins, S. S. (1962). *Affect, imagery, consciousness. Volume I. The positive affects*. New York: Springer.

Tribus, M. (1961). *Thermostatics and thermodynamics*. Princeton, NJ: van Nostrand.

van Ditmarsch, H. P., van der Hoek, W., & Kooi, B. P. (2007). Dynamic epistemic logic. *Synthese Library*, 337. Berlin: Springer.

Weiner, B. (1985). "Spontaneous" causal thinking. *Psychological Bulletin, 97*, 74-84.

Wooldridge, M. (2002). *An introduction to multiagent systems*. West Sussex: John Wiley & Sons.

Wundt, W. (1863). *Vorlesungen über die Menschen- und Tierseele [Lectures on the mind of man and animals]*. Leipzig: Voss.

KEY TERMS

Affective: Colloquially: concerned with or arousing feelings or emotions; emotional. In today's psychology, "affective" is often used as a cover term for all emotional and related phenomena (emotions, moods, evaluations...).

Agent(s): An autonomous entity capable of action.

Anticipation: In humans, "anticipation" refers to the mental act or process of "looking forward" by means of forming predictions or beliefs about the future. An anticipatory agent is a natural or artificial agent who makes decisions based on predictions, expectations, or beliefs about the future.

Artificial Surprise: Surprise synthetized in machines (artificial agents), usually intended as a simulation of surprise in natural agents, specifically humans. Depending on context, "surprise" may either refer to the mechanism that produces surprise, or to its product, the surprise generated.

Astonishment: A subform of surprise distinguished from regular surprise, according to different authors, by higher intensity, longer duration, or special causes (e.g., fully unexpected events [astonishment] in contrast to misexpected events [ordinary surprise]).

Belief: In humans: a mental state (propositional attitude) in which a person holds a particular proposition *p* to be true. In artificial agents: a corresponding functional (processing) state.

Computational Model(s): A computational model is a computer program that attempts to simulate a particular natural system or subsystem.

Conflict(s): See "mismatch."

Disappointment: The unpleasant feeling resulting from an expectation failure concerning a desired event, or put alternatively, the disconfirmation of the belief that the desired event would occur.

Emotions: In humans: mental states subjectively experienced as (typically) positive or negative feelings that are usually directed toward a specific object, and more or less frequently accompanied by physiological arousal, expressive reactions, or emotional behaviors. Typical examples are joy, sadness, fear, hope, anger, pity, pride, and envy. In artificial agents: corresponding processing states intended to simulate emotions of natural agents, usually humans. Note that depending on context, 'emotion' may also refer to the mechanism that produces emotions rather than to its products.

Expectation: In common parlance, an expectation is a belief regarding a future state of affairs. In the literature on surprise, "expectation" is frequently used synonymously with "belief".

Unexpected: A proposition *p* is unexpected for an agent *A* if *p* was explicitly or implicitly considered unlikely or improbable to be true by *A*, but is now regarded as true by *A*.

Mismatch: Discrepancy or conflict between objects, in particular a contradiction between propositions or beliefs.

Misexpected: A proposition *p* is *misexpected* for an agent *A* if *p* is detected by *A* (or a subsystem of *A)* to conflict with, or to mismatch, a pre-existing, specific and usually explicit belief of *A* regarding *p*. In contrast, *p* is *unexpected* for *A* in the narrow sense of the word if *p* is detected by *A* to be inconsistent with *A*'s background beliefs. Finally, *p* is unexpected for *A* in the wide sense of the term if *p* is either misexpected for *A*, or unexpected in the narrow sense.

Surprise: In humans: a peculiar state of mind caused by unexpected events, or proximally the detection of a contradiction or conflict between newly acquired and pre-existing beliefs. In artificial agents: a corresponding processing state caused by the detection of a contradiction between input information and pre-existing information. Note that depending on context, "surprise" may also refer to the mechanism that produces surprise, rather than to its product.

ENDNOTE

[1] The assumption that surprise is elicited by unexpected events (events that disconfirm an explicit or an implicit expectancy or belief) is made by practically all classical and modern surprise theorists, and also is part of common-sense psychology (Reisenzein, 2000a; Ruffman & Keenan, 1996). However, there is some variation in how this assumption is worked out (see e.g. Charlesworth, 1969; Desai, 1939; Ortony & Partridge, 1987; Shand, 1914; and the computational models of surprise described in the next section). Note also that, whereas in everyday language, expectations are a subspecies of beliefs (namely, beliefs that refer to future states of affairs), in the technical literature reviewed here, "expectation" is usually used as a synonym of "belief". Because one can also be surprised about past and atemporal states of affairs, this broad reading of "expectation" is needed in discussions of surprise.