

Copyright © 1987. International Joint Conferences
on Artificial Intelligence, Inc.
All rights reserved.

Edited by
John McDermott

Distributed by
Morgan Kaufmann Publishers, Inc.
95 First Street
Los Altos, California 94022

ISBN 0-934613-43-5

Printed in the Netherlands
through Interprint, San Francisco.

SURPRISINGNESS AND EXPECTATION FAILURE: WHAT'S THE DIFFERENCE?

Andrew Ortony
University of Illinois at Urbana-Champaign
and
Derek Partridge*
University of Exeter, UK

Intelligent systems operate in the midst of a superabundance of information lacking the tags that indicate which few aspects are significant to the particular problems at hand at any given time and place. Given this wealth of information coupled with real-time processing constraints, selective attention is fundamental to any chance of success. In much of cognitive science, attentional focus is linked with the concept of expectation generation and failure. Unfortunately, research in AI has to finesse the problem of selective attention, although sometimes recognizing it as an important component of the frame problem (e.g., Schank, 1979; 1982).

We think that surprisingness, which is closely related to expectations and expectation failures, is important for focusing attention. However, we shall argue that there is much more to surprisingness than expectation failure, and that the difference between the two is critical for AI. The standard view of the role of expectations in AI suffers from a reliance on underspecified concepts of "expectation" and "expectation failure," both of which appear to be oversimplifications within which a range of rather different phenomena are confounded. We shall attempt to provide a more general preliminary analysis of surprisingness.

To start, we need to introduce some constructs that we use in our analysis. First, we shall suppose that a system is presented with an *input proposition*. Second, we shall assume that at any point in time the current task results in some parts of the data base being *activated*. Third, we assume that the data base comprises both episodic and semantic knowledge, an important aspect of which is that each propositionally distinct element has an *immutability index* associated with it. In other words, some elements are believed (perhaps erroneously) to be virtually immutable, while others are believed to be typically true, and yet others merely sometimes true. Whereas we view immutability as a continuous variable, for simplicity of exposition it will be sufficient to distinguish cases that are *immutable* from those that are merely *typical*. This is because when surprisingness results from a conflict between an input proposition and propositions explicitly represented or readily deducible from the data base, there has to be a reasonable degree of conflict. It is generally surprising to encounter situations that conflict with propositions that are held to be *immutable*, but it is not generally surprising to encounter propositions that conflict with things that are only sometimes true. Sometimes dogs are black, but this fact alone does not give rise to surprise on encountering a brown dog.

We assume that with the help of a few simple inference rules, the knowledge base has the potential to produce a wealth of knowledge not explicitly represented. The set of propositions that are explicitly represented in the data base, or that are readily deducible from those that are, we call the *practically deducible* propositions. Thus, the class of practically deducible propositions includes explicitly represented propositions as limiting cases. Practically deducible propositions are contrasted with what we call

practically non-deducible propositions. These are possible propositions that are neither explicitly represented nor readily deducible from those that are. Not all propositions that are formally deducible count as practically deducible. In general, we view a proposition as being practically deducible to the extent that it does not require many and complex inferences. Examples of practically deducible propositions might include propositions to the effect that (a) restaurants are establishments in which meals are served (which might be explicitly represented and virtually immutable), or (b) that in restaurants one pays *after* eating (which might be deducible from a script and represented as being typically true), or (c) that 628 is not a prime number (which is deducible and immutable). A non-deducible proposition might be that a rock will come hurtling through one's office window in five minutes, or indeed, the proposition that a rock will not come hurtling through one's window in five minutes.

Although a proposition may be practically deducible, it does not follow that it will in fact have been deduced on any particular occasion in which it is implicated in some reaction of surprise. In principle, it might have been, but in practice it may not have been. When it has been deduced, we can say that the corresponding event is *actively expected*, in which case we have what might be called an *active prediction*. When it has not, one might refer to the situation as one of *passive expectation*, or *passive assumption*. This amounts to proposing that expectations (active or passive) can only pertain to *practically deducible* propositions. So, for example, if one actively expects that the Democratic candidate will win the next presidential election, this prediction is rooted in propositions such as that there is to be an election, and that there is a Democratic candidate. If the Republican candidate wins, one would be surprised as a result of the failure of an active expectation (in the sense that the expected outcome was, at some point, actively inferred as a prediction). However, if, while having the same data base, one had never made the inference to the conclusion of an expected Democratic victory, the surprise would be the result of a *passive* expectation failure. Such failures are better thought of as failures of assumptions in that they derive from input propositions that are inconsistent with what one knows or believes but that were not actively predicted. Some deviations from normalcy (Kahneman & Miller, 1986) should probably also be included in this category. There remains, however, an important third source of surprise, namely cases that do not arise from conflicts at all. In such cases the surprise results not from an input proposition violating an (active) expectation or a (passive) assumption, but from the input proposition being not practically deducible from the data base. In such cases there is no logical contradiction between the input proposition and a practically deducible proposition. Rather, there is a conflict between the input proposition and what, after the fact, may be judged to be normal or usual. Such retrospective judgments of unexpectedness are the subject of Kahneman and Miller's (1986) Norm theory.

So we have the following important distinctions: First, states of the world corresponding to practically deducible propositions are in principle *predictable*, while states of the world corresponding to practically non-deducible propositions are not.

* This paper was prepared while Derek Partridge was at the Computing Research Laboratory of New Mexico State University whose support we gratefully acknowledge.

Second, for states of the world corresponding to deducible propositions, there can be expectations if the related propositions were activated. These expectations are *active* (predictions) if they have been consciously entertained, and they are *passive* (assumptions) if they have not. Third, for states of the world corresponding to non-deducible propositions, there cannot be expectations (active or passive), but only *ex post facto* evaluations of surprisingness on the basis of deviations from norms. It is our suspicion that in the real world, a great many surprising situations are of this latter variety, involving events whose prospects were never entertained but that just arose "out of the blue." It is for this reason that one needs to be cautious about equating surprisingness with expectation failures.

The Table below presents the major sources of surprise that this kind of analysis yields.

STATUS OF INPUT PROPOSITION	NATURE OF AFFECTED PROPOSITIONS	RELATED COGNITION	
		ACTIVE	PASSIVE
DEDUCIBLE	IMMUTABLE	[1] $S_A = 1$ prediction	[2] $S_P = 1$ assumption
	TYPICAL	[3] $0 < S_A < 1$ prediction	[4] $S_P < S_A$ assumption
NOT DEDUCIBLE	IMMUTABLE	[6] no entry	[8] $S_P = 1$ none
	TYPICAL	[7] no entry	[9] $0 < S_P < 1$ none

Consider first what happens when an input proposition conflicts with a (deducible) active prediction or a passive assumption. Recall that an active expectation is one that is being or has been consciously entertained. Such a conflict can legitimately be called an *expectation failure* in that there was an explicit expectation (rooted in a prediction) that failed. An example might be a case in which one chooses to go to a French restaurant because one feels like eating French food. On receiving the menu, one discovers that all the entrees are Greek. Such a situation would be highly surprising because it violates a virtually immutable constituent of the French restaurant schema which was active in the sense that it was a causal component of a decision to choose the restaurant. This source of surprisingness is shown in the Table in cell [1], in which the degree of surprisingness is represented by S_A , where the subscript, A, indicates that the surprisingness results from an Active prediction. In this particular case, $S_A = 1$, indicating a maximum degree of surprise. In fact, however, the surprisingness of such an event is independent of whether the violated expectation was active (i.e., predicted) or passive (i.e., merely assumed). The case of a violated assumption is shown in cell [2], where the surprisingness, S_P , is written with a subscript, P, to indicate that it results from a Passive expectation. An example would be a case in which one came across a green dog. Given a reasonable representation of DOG (in which, of course, the proposition that dogs are not green would not be explicitly represented), it is not difficult to infer that dogs are not green. However, in the normal course of events, such an inference would not be made although, to the degree that it can be readily inferred, it is a practically deducible proposition. Thus, an input proposition to the effect that there is a green dog would conflict with a passive expectation for dogs not to be green. The high degree of surprise indicated in cells [1] and [2] suggests that the surprise results from the fact that it was an *immutable* proposition that was violated.

In the case in which the violated proposition is only *typically* true as opposed to immutable, the level of surprise tends to be somewhat less high. For example, if, having chosen a French restaurant because one wanted to eat frogs' legs, the waiter explains that frogs legs are not a menu item in this particular restaurant, this might be somewhat surprising because typically (but by hypothesis, not necessarily) frogs legs can be ordered in French restaurants. This would be an example of the failure of an active expectation of something that is typically true. In such a case we might suppose $0 < S_A < 1$ (cell [3]). What this indicates is that violations of propositions that are only typically true are less surprising than violations of immutable ones, which are at the limit of surprisingness. In fact, the same analysis applies to failures of passive expectations (i.e., assumptions) about things that are typically true, although here we indicate a slightly reduced level of surprise because the expectation was not an active prediction (cell [4]).

We move now to a discussion of cases in which surprise results from non-deducible propositions. There are two general classes of cases here, rather than the four in the practically deducible cases. This is because, by definition, one cannot actively predict something that is not realistically derivable from the data base, which means that the input proposition cannot, in principle, conflict with an active expectation. This is why there is no entry in cells [5] and [7]. The kind of examples we have in mind in connection with input propositions that are not practically deducible are ones in which something happens that was not and could not (realistically) have been entertained. When a rock flies through one's office window, one will certainly be surprised, but the surprise cannot be attributed to a conflict with some practically deducible proposition. This is the case numbered [8] in the Table. What we are claiming here is that the proposition that a rock will not fly through one's office window is not practically deducible. Before we elaborate this example, we need to make a couple of preliminary observations. First, we are assuming a normal context, so that we would not want to make this claim for a person who looks out of his window and sees a riot taking place, with people hurling rocks in all directions. In such a situation, it would be quite possible to entertain the possibility. The proposition would then be practically deducible because new and relevant propositions would have been added to the knowledge-base. Second, as mentioned earlier in explaining what counts as a practically deducible proposition, in normal contexts, the connection between what one knows and propositions about rocks flying through windows is so remote that it cannot be deduced. Furthermore, there is nothing to *motivate* an attempt to deduce it. It is not realistic to suppose that a system with goals and tasks to perform is at the same time randomly spawning inferences about unrelated and improbable possibilities. This is what distinguishes the case of the rock from that of, say, tossing a coin. When a coin is tossed, the propositions that it might come up heads and that it might come up tails are (presumably) generated with more or less equal confidence, and their generation is motivated by the context.

So, when the input proposition is a practically non-deducible proposition it cannot conflict with an expectation. It can, however, be *inconsistent* with practically deducible propositions that it *suggests*, and this inconsistency might be with relatively immutable propositions or only with ones that are typically true. If it is inconsistent with relatively immutable propositions the situation is very surprising (shown as $S_P = 1$ in cell [8]). Such a case might arise were one to suddenly see a person take off and fly with no apparent mechanical aids (Superman style). If, on the other hand, the conflict is with a typical rather than an immutable proposition such as finding a restaurant in which one does not sit down to eat, or having a rock fly through one's window, we have the kind of surprisingness indicated in cell [9].

Schank (1982) offers an account of learning and reminding in terms of expectation failures that derive from predictions encoded in scripts. An active script generates active expectations,

all other scripted propositions (presumably) account for (what we term) passive expectations. Schank briefly acknowledges the difficulties of, what for us are, type 8 expectation failures, but only to say that they are not "failed expectations in the straightforward sense of that term". Thus, when Schank talks about "expectations" he is referring to what we have identified as active expectations. Yet, if one restricts expectation failures to the failures of active expectations, and postulates nothing else, it is difficult to avoid the conclusion that there can be situations in which one cannot possibly be surprised even though one has no idea what to expect. The solution to this problem is to allow that input propositions that are practically non-deducible can give rise to surprise. The omission of this class of sources of surprise (cells [6] and [8] of our Table) from Schank's discussion is probably the result of focusing too closely on the scripts and their capacity to generate active expectations at the expense of a more global view of the problem of expectations and surprise. Surprisingness resulting from a nondeducible input proposition cannot, by definition, be accounted for in terms of script conflict.

The term "expectations violation" is used by Hayes-Roth (1983) to label a component of his description of "five heuristic methods for repairing flawed beliefs". The "beliefs" described are strategies for playing a card game, and his usage of the "expectations violation" concept is minimal. Input data may disconfirm a prediction from a strategy for playing the card game. He considers only the case in which an expectation, that is, a prediction, turns out to be wrong (cell [3] in our Table).

Partridge (1985) has explored the scope and various manifestations of an input-expectation discrepancy reduction mechanism. The numerous examples discussed illustrate the importance of this mechanism at a number of different levels of interpretation of intelligent behavior: cognitive, subsymbolic, and neuronal. The examples are also used to expose the range of differing limited perspectives from which input-expectation discrepancy is viewed. The paper argues for the importance of this discrepancy both as an attention-focusing mechanism and as the instigator of learning behaviors (with some, but by no means total, guidance as to what needs to be learned).

Dennett (1984) notes that a midnight-snack-making robot should be surprised by the fact that its beer glass has been glued to a shelf. He observes that a startle response means we must have expected something else, but that things are not that simple. One can be startled by something one didn't expect without having to expect something else. In terms of our analysis, the glued-down beer glass is an example of a cell [8] situation in which S_p is perhaps closer to 1 than to 0. The robot's expectation that the glass would not be glued to the shelf is not a practically deducible one, and glasses are not typically glued to shelves but there is no reason why one could not be so glued. Our contention that for this example S_p will be at the high end of its range is due to the fact that the glued-down glass directly thwarts the robot's active plan to pick up the glass. By way of contrast, in the absence of some exceptional circumstance, the occurrence of say, alphabetically parked cars (e.g. an Alpha Romeo, a Buick, and a Chevrolet parked in line) again falls into cell [8] of our table. But this time S_p will be at the low end of its range. It may even be zero if the mechanism of selective attention fails to register either the parked cars themselves or the alphabeticity of the parked threesome. Dennett appears to have purposely selected cases characterized by cells [6] and [8] in our Table, presumably precisely because such examples are particularly difficult to explain under any simplistic interpretation of surprise as expectation failure.

As is clear from our introductory remarks, we believe that the registration of and reaction to failed expectations, failed assumptions, and unanticipated incongruities, are a crucial component of general intelligence. These sources of surprisingness can provide the basis for an attention-focusing mechanism, and a cue for learning. They are nothing like a complete answer to either problem, but they do constitute an important piece of the

solution, a piece that has yet to be carefully explored in AI. Thus we would expect to find considerations of surprisingness in a problem solver that is not pre-set to solve only one very specific problem, and as a precursor to machine learning where the raw input data are not highly preprocessed to remove all information that is irrelevant to the specific rule or concept being learned:

- Dennett, D. (1984). *Cognitive Wheels: the frame problem of AI*. In O. Hookway (Ed.), *Minds, Machines and Evolution*. New York: Cambridge University Press.
- Hayes-Roth, F. (1983) Using proofs and refutations to learn from experience. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning*. Palo Alto, CA: Tioga.
- Kahneman, D. & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.
- Partridge, D. (1985) Input-expectation discrepancy reduction: A ubiquitous mechanism, *Proc. 9th IJCAI*, Los Angeles, CA, 267-273.
- Schank, R. C. (1979). Interestingness: Controlling Inferences. *Artificial Intelligence*, 12, 109-117.
- Schank, R. C. (1982). *Dynamic memory*. New York: Cambridge University Press.