TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 11 (2019) 103–118

ISSN:1756-8765 online
DOI: 10.1111/tops.12345

This article is part of the topic “The Ubiquity of Surprise: Developments in Theory, Converging Evidence, and Implications for Cognition,” Edward Munnich, Meadhbh Foster and Mark Keane (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

Seeing Patterns in Randomness: A Computational Model of Surprise

Phil Maguire,[a] Philippe Moser,[a] Rebecca Maguire,[b] Mark T. Keane[c]

[a]*Department of Computer Science, Maynooth University*
[b]*Department of Psychology, Maynooth University*
[c]*School of Computer Science and Informatics, University College Dublin*

Received 28 June 2016; received in revised form 5 December 2017; accepted 27 February 2018

Abstract

While seemingly a ubiquitous cognitive process, the precise definition and function of surprise remains elusive. Surprise is often conceptualized as being related to improbability or to contrasts with higher probability expectations. In contrast to this probabilistic view, we argue that surprising observations are those that undermine an existing model, implying an alternative causal origin. Surprises are not merely improbable events; instead, they indicate a breakdown in the model being used to quantify probability. We suggest that the heuristic people rely on to detect such anomalous events is *randomness deficiency*. Specifically, people experience surprise when they identify patterns where their model implies there should only be random noise. Using algorithmic information theory, we present a novel computational theory which formalizes this notion of surprise as randomness deficiency. We also present empirical evidence that people respond to randomness deficiency in their environment and use it to adjust their beliefs about the causal origins of events. The connection between this pattern-detection view of surprise and the literature on learning and interestingness is discussed.

Keywords: Surprise; Randomness deficiency; Algorithmic information theory; Bayesian reasoning; Stochastic model; Representational updating; Interestingness; Data compression

Correspondence should be sent to Phil Maguire, Department of Computer Science, Eolas Building, North Campus, Maynooth University, Maynooth, Co. Kildare, Ireland. E-mail: pmaguire@cs.nuim.ie

[Correction updated on June 23, 2018 after first online publication: The affiliation of the author Mark T. Keane has been changed from “University of California, Davis” to “University College Dublin”.]

As we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know.

—Donald Rumsfeld, February 12, 2002, United States Secretary of Defense

1. Introduction

Every day, people deploy their knowledge to carry out a bewildering array of cognitive tasks, ranging from motor tasks to perception, to reasoning and decision making. This knowledge, as highlighted by Rumsfeld above, is tenuous. Not only are facts themselves uncertain, but so too are the assumed causal models that allow these facts to be inferred in the first place. At any moment, we may be faced by a revelation which necessitates a fundamental reevaluation of our beliefs, resulting in an experience we know as "surprise."

Given the prevalence of uncertainty in the world, surprise is a ubiquitous experience. It can be elicited by many different events, including unique or unusual occurrences (e.g., finding a large wad of money on the ground; meeting your neighbor while on holiday in a foreign country), to more mundane events (e.g., realizing your keys are not where you initially thought; finding that there is no milk in the fridge). While numerous accounts have been proposed to explain the experience of surprise (Baldi & Itti, 2010; Maguire, Maguire, & Keane, 2011; Meyer, Reisenzein, & Schützwohl, 1997), debate persists regarding the precise factors that contribute to the perceived surprisingness of an event (Foster & Keane, 2015).

Because a theory of surprise must cover the failure of *every* possible model of reality, it must be powerful and general. It must be capable of discriminating between observations that are typical and unthreatening versus those that signal a breakdown in modeling. Whatever this mental mechanism is, it cannot just be an ordinary theory that itself is subject to failure; it must be a fail-safe meta-theory, grounded on deep inviolable principles. In this article, our goal is to explain and formalize a model of surprise that meets these requirements. To do this, we will first review insights from classical and Bayesian probability theory, before outlining our own computational model of surprise based on the principles of algorithmic information theory (AIT). Specifically, we demonstrate how the concept of randomness deficiency can be relied on to successfully model surprise.

2. Probability theory and surprise

Over the last century, probability theory has emerged as the ubiquitous approach for modeling uncertainty and surprise, though its relatively late development (17th century) hints that it is not a particularly natural way of thinking. Probability theory was formalized by Kolmogorov in the 1930s through the notion of probability space, whereby a set of possible outcomes is mapped to a number that represents its likelihood by a probability measure function (see Li & Vitányi, 2008). It has been argued that surprise holds an inverse relationship with probability, in that highly probable outcomes result in low levels of surprise and vice versa (e.g., Meyer et al., 1997). However, the concept of a direct relationship between the two does not always hold up to scrutiny (Maguire & Maguire, 2009).

A key feature of classical probabilistic models is that their outcomes are assumed to be independent of each other, and hence independent of the model itself. As such, outcomes do not provide any information about the model, nor do they assist in predicting subsequent outcomes. Probability theory works perfectly for describing the behavior of finely calibrated stochastic mechanisms, such as dice. Each roll of a dice can be considered as independent, insofar as subsequent rolls do not cause observers to update their beliefs about the dice.

However, since real-world evidence in support of any stochastic model is necessarily finite, classical probabilistic models, which do not include any role for learning, are vulnerable to being undermined by certain possible, though somehow “atypical,” outcomes. For instance, if a dice produced a six time after time, eventually observers would find this “surprising.” Repeated observations of the same number undermine the assumption of independent rolls, and they lead observers to question whether the dice might be biased.

Although a repeated sequence of sixes is mathematically just as likely as any other particular sequence, it is somehow unsatisfactory as a “random” sequence. This idea was demonstrated in practice in 2009, when the same set of six numbers appeared in two consecutive draws in the Bulgarian National Lottery. While lottery officials insisted that manipulation was impossible, a commission was nevertheless established to investigate the incident, indicating a lack of complete confidence in the randomness of the draw. Maybe the balls were not equally weighted, maybe the drum mechanism was defective, or maybe the lottery officials were corrupt. In sum, the occurrence of the same set of numbers on two consecutive occasions was so surprising that it served to undermine confidence in the assumption of independent draws.

This example illustrates that equally probable events are not always perceived as equally surprising (see also Teigen & Keren, 2003). Moreover, it demonstrates that supposedly strong assumptions, such as the randomness of a lottery draw, can be challenged by only brief deviations from expected randomness. In practice, people have to be ready to alter their beliefs at all times, meaning they need to be alert to signals that something is amiss in their representation of the environment. Adherence to a fixed stochastic model precludes any possibility of learning, a failure which would be detrimental to survival if applied in the real world (Baldi & Itti, 2010). Because it is not possible to be absolutely certain of anything, people need to ready for surprises.

In the following sections, we seek to define the property that highlights certain observations as being surprising, and to show how people rely on this signal to update their representations in response to such anomalies. We begin with the Bayesian theory of surprise presented by Baldi and Itti (2010), before revealing our own computational theory.

3. Bayesian surprise

Whereas, classical probability is founded on certainties and fixed causal models, Bayesian probability instead focuses on subjective uncertainties and inductive inference. It extends the classical approach by viewing probability not as an objective phenomenon, but as the state of knowledge of an uncertain observer, a state which is therefore subject to refinement. Bayesian probability opens up a new dimension to classical probability, in that it takes into account how observations *inform* the observer. It is an approach that has proved useful for psychological modeling, having been applied to the understanding of a wide range of cognitive phenomena (see Chater, Oaksford, Hahn, & Heit, 2010). The theory recasts many specific cognitive tasks in the form of a universal framework whereby predictions, made from prior hypotheses, are subsequently updated on the basis of observed evidence. Bayesian graphical models, for instance, can explicitly express which pieces of evidence are dependent on others (Pearl, 2009; Sloman & Lagnado, 2005), and it can predict interesting patterns of inference in people's reasoning (Rottman & Hastie, 2016).

An influential Bayesian explanation of surprise has been proposed by Baldi and Itti (2010). Here, surprise is deemed to be a product of the amount of learning that has taken place by an observer, which is quantified in terms of the relative entropy (i.e., Kullback–Leibler divergence) between the prior and posterior distributions. To quantify the surprisingness of an observation, one needs to identify a set of hypotheses which are being actively considered. The surprise value reflects the extent to which the balance of belief gets shifted within that hypothesis set.

Experimental results have reinforced the value of Baldi and Itti's (2010) surprise notion. It has been shown to yield a robust performance in predicting human gaze across different spatio-temporal scales, modalities, and levels of abstraction (Friston et al., 2012; Itti & Baldi, 2006), as well as being applied to detect salient acoustic events (Schauerte & Stiefelhagen, 2013). Other theoretical work, which views surprise in terms of contrasts between hypotheses (e.g., Meyer et al., 1997; Teigen & Keren, 2003), is broadly consistent with Baldi and Itti's (2010) approach.

This account, however, does not seem to cover all instances of surprise, especially those that are completely unanticipated. For example, if a brick comes through the window, it is surprising precisely because one would never have thought of entertaining such a hypothesis (Ortony & Partridge, 1987). There are also examples of surprising experiences that are awkward to express in terms of belief shifting among a finite number of discernible alternatives, such as looking down at the dashboard of a car and seeing the odometer at 66,666 km (Dessalles, 2008).

As stated by Baldi and Itti (2010, pp. 661–662) themselves: "The process by which a learning system realizes that a model class is unsatisfactory in an alternative free setting—the open-ended aspect of inference—has so far eluded precise formalizations and ought to be the object of future investigations." Baldi and Itti's (2010) theory might therefore be more accurately regarded as a theory of "surprise calculation" (i.e., informativeness) rather than as a theory of "surprise detection." Rather than capturing that initial "oh-no" moment of heightened awareness before a stimulus is fully explained, it instead provides a final retrospective evaluation of how much impact an event has had in shifting a set of beliefs.

In contrast, Foster and Keane (2015) view surprise as a metacognitive estimate of the work involved in explaining an abnormal event, an estimate which becomes available before any belief shifting has even been attempted. An anomalous observation, which initially provokes a high metacognitive estimate of discrepancy, may in the end turn out not to justify any representational updating at all. For example, if you were to look up at the sky and see a cloud in the shape of a dog, a high level of surprise detection would be experienced. On reflection, however, any causal account might seem unjustified, leading you to view the event as a “fluke” rather than updating your beliefs. Following Baldi and Itti’s (2010) formulation, which is based on ultimate belief adjustment, the event would register low on surprise calculation. In sum, Baldi and Itti’s approach fails to explain why a striking coincidence without immediate explanation should grab people’s attention.

A comprehensive theory of surprise needs to account for how people flag observations as being potentially anomalous before they’ve figured out how to make sense of them: People have to be able to detect surprise before resolving surprise. In the following section, we propose a more general “hypothesis-neutral” theory of surprise, which does not depend on quantifying Bayesian learning.

4. Surprise as randomness deficiency

We propose that surprising observations are those that contain unaccounted-for patterns; in other words, they evidence *randomness deficiency*. When a set of observations contains patterns where an existing model implies there should only be random noise (e.g., a dice continually landing on 6), it undermines the assumption of independent outcomes and suggests the presence of some alternative mechanism at work. If the deviations of a model’s predictions are randomness deficient, it implies that a superior predictive model is available (see Vitányi & Li, 2000). The experience of physiological surprise can be viewed as a response to this situation, whereby heightened awareness and enhanced sensory intake (e.g., eye widening, opening of the mouth, enlargement of the nasal cavity serve) serve to facilitate the resolution of the discrepancy (Susskind et al., 2008; Tottenham et al., 2009).

To illustrate the concept of randomness deficiency more clearly, consider the situation where a packet of rice is accidentally spilled on the floor. According to probability theory, it is just as likely that the grains of rice will produce a “pattern” as any other particular arrangement. Nevertheless, we expect the grains to scatter “randomly.” What that means in practice is that we do not expect to see any out-of-context patterns arise in the arrangement of the grains. Imagine then how surprised you would be to find your name spelled out precisely by the rice (in 1795 Pierre-Simon Laplace used a similar example of finding words spelled out by random letters; see Griffiths & Tenenbaum, 2007).

The rice-name scenario is randomness-deficient because the outcome can be concisely described, suggesting the potential for another explanation that provides a better fit: Perhaps somebody wrote your name on the floor in a sticky substance that has attracted the grains. Adjusting your understanding in this way may be more reasonable than accepting that a random scattering of rice grains would just happen to spell out your name.

Rather than holding an innate appreciation of what randomness should look like, people are instead attuned to detecting the presence of patterns (Zhao, Hahn, & Osherson, 2014). An observation appears random when there is no obvious way to describe it more concisely than using the assumed stochastic model and recording each event independently. For example, given a random sequence of dice rolls, the most concise way to represent it is simply to write down each roll result separately, one after the other. In contrast, an observation is randomness deficient (i.e., surprising) when it can be described more concisely than this; in such cases, we say that the sequence can be "compressed." For example, while an apparently random sequence such as 8, 11, 20, 29, 31, and 45 cannot seemingly be described in any more concise fashion, a sequence such as 1, 2, 3, 4, 5, and 6 can be compressed as "1 to 6," suggesting it is not the product of random selection.

The greater the compressibility of a supposedly random set of observations, the greater its randomness deficiency, and the greater its likelihood of having being produced by an alternative mechanism (see Vitányi & Li, 2000). Sequences that can be described by simple patterns are extremely rare for a random lottery draw, yet presumably much more common for other deterministic mechanisms. Consequently, if we observe a lottery sequence with a pattern, it surprises us, because it leads us to question whether the draw might be biased.

To define a general theory of surprise, it is necessary to formalize this concept of randomness deficiency. In the following section, we show how AIT, the discipline within which the concept of randomness has been defined, can provide an appropriate framework. Our idea is simply this: Surprise is the normalized difference between the probabilistic point of view, which treats observations as independent, and the computational point of view, which gives the shortest possible description. The more the probabilistic encoding deviates from the shorter computational encoding, the greater the level of surprise.

4.1. Preliminaries

The fundamental premise of AIT concerns the equivalence between likelihood and simplicity (Chater, 1996; Chater & Vitányi, 2003). Specifically, AIT is based on the assumption that the probability of a model, grammar, or pattern, is inversely proportionate to its complexity (complexity being the opposite of simplicity), and that complexity is proportionate to minimum description length (Chater & Brown, 2008). In other words, the more concisely a model describes a set of data, the more likely it is to be correct, and the more successful its predictions will be. This finding can be interpreted as a formalization of Occam's razor, the idea that, all being equal, simple theories should be preferred because they are more likely (Chater & Brown, 2008). Vitányi and Li (2000) show that data compression (i.e., looking for models that support concise descriptions of events) is almost always the best strategy, both in hypothesis identification and prediction.

Kolmogorov complexity is a notion which captures the most concise possible description of a dataset, thus allowing us to define the computational view (see Li & Vitányi, 2008, for more details). The main idea is as follows: Suppose we flip a coin 30 times to generate a random binary string of length 30. We obtain the string 100100010011010011011111111100 and are happy with the result, since it looks "random." The next day, we repeat the experiment and obtain the string 000000000000000000000000000000. This time we are unsatisfied, since the string does not look random at all. How can we formally argue that the first string looks random, whereas the second does not? We cannot argue from the probabilistic view because the two strings are equally probable to be the result of 30 coin flips (both have probability 2^{-30}). So let us try instead to describe the first string as succinctly as we can. Since we can spot no obvious pattern, the shortest description we can come up with is: "Print 100100010011010011011111111100." What about the second string? In that case we find a much shorter description, namely: "Print thirty zeros." So, while the first string admits no short descriptions (the length of the shortest description we could find is as long as the string itself), the string of 30 zeros admits a description of length 18 (namely "Print thirty zeros"), which is much shorter than 30. Herein lies Kolmogorov's idea: the first string is random because the length of its shortest description is as long as the string itself, while the second string is not random, because its shortest description (i.e., its complexity) is substantially shorter than the length of the string.

The idea is very elegant, but what language should we use to describe strings? Should we use English? Latin? Esperanto? What if in some ancient language the short word "chisen" stands for "100100010011010011011111111100"? Kolmogorov observed that it does not matter as long as one uses a fixed "universal" description language based on a universal computer programming language: Since one quickly runs out of short descriptions for long strings, any two universal languages will yield almost the same set of random strings (with the exception of only a few strings). To allow mathematical formalization, Kolmogorov describes strings based on the idea of a universal computer, that is, a computer that can simulate any other computer (formally a universal Turing machine). Kolmogorov proved that the choice of the universal computer does not matter: If one chooses another universal computer, then the lengths of the shortest descriptions based on the new computer differ from the ones based on the old universal computer by at most a fixed additive constant. Also, we can restrict ourselves to binary strings, since any finite object (English sentences, books, movies, etc.) can be encoded into such strings in some natural way.

Accordingly, a valid description of a string x is the code of a program p, such that if we run the program p on the universal computer, the computer prints x. The Kolmogorov complexity of string x is the length of the shortest program p that makes the universal computer output x. In this case, the shortest program for the string 100100010011010011011111111100 is at least 30 bits long, while the string of 30 zeros has a lower Kolmogorov complexity, because it can be computed by a short program.

We say a string is random if its Kolmogorov complexity is greater than or equal to its length. Otherwise we say the string is not random.

4.2. Formal definition of surprise

Imagine a black box that prints out strings, one after the next. We do not know how it works, but for every possible string x we know the probability $\Pr_B(x)$ that the black box outputs x (formally $\Pr_B$ is a probability distribution over the set of binary strings).

Knowing the distribution $\Pr_B$, there are some "typical" strings we expect to be outputted, whereas some others are *surprising* (the atypical ones). One can quantify this lack of typicality in terms of randomness deficiency (see Li & Vitányi, 2008). For example, consider box B that outputs all the binary strings of length 30, where each such string is outputted with probability 2^{-30}. Imagine the box outputs string s = 100100010011010011011111111100. We are not surprised. Let us compute the randomness deficiency of s: Suppose we want to encode the outputs of the black box using a probability based encoding which assumes all outputs are independent of each other. Then s would require $\log 1/\Pr_B(s) = 30$ bits to encode. What if we use the universal computer instead? Imagine for fairness we provide the computer with the probability distribution $\Pr_B$ (i.e., the computer can look up the value of $\Pr_B(x)$ for any string it wishes, and it is given the answer; this ensures that the length of the computational encoding is never greater than the probabilistic one). Then the shortest description of s has length approximately 30 (knowing $\Pr_B$ does not help the computer in this case), and the difference between the probabilistic and computational views (i.e., the randomness deficiency) is approximately zero.

Imagine the box next outputs t = 000000000000000000000000000000. Why does this seem surprising? Let us investigate by computing the randomness deficiency of t: The probability based description length is still 30. If the shortest description of t on the universal computer given $\Pr_B$ is 18, then the randomness deficiency is $30 - 18$, which is much larger than for the other string s.

In sum, given a black box with probability distribution $\Pr_B$, we propose that the surprisingness of string x is expressed by the randomness deficiency of x, that is, the difference between the length of the probability based encoding of x and the length of the shortest description of x on the universal computer given $\Pr_B$. This can be written formally as $\delta(x|\Pr_B) = \log 1/\Pr_B(x) - C\,(x|\Pr_B)$.

This value can be normalized to a value between 0 and 1 (up to a logarithmic factor), thereby quantifying bits of surprise per bit of observation, in other words, the proportion of the probability based encoding that is superfluous according to the computational point of view. The more an observation can be compressed (i.e., the greater the discrepancy between the probabilistic and computational points of view), the greater the associated level of surprise.

4.3. Discussion

Randomness deficiency applies to situations where we expect randomness but then experience structure. It also applies to situations where we expect structure but experience randomness instead (see Loewenstein & Heath, 2009, for an account of how such pattern breakers are used to generate surprise and amusement in folktales and story jokes). Imagine, for example, having observed the following sequence: 1, 2, 3, 4, 5, 6, 7, ... Clearly, one would be more surprised if the following number was 86 than if it was 8, even though 86 is ostensibly more random. What differs between this case and the lottery example is the probability distribution. Assuming people have adjusted their representation so as to anticipate an ascending sequence, the appearance of 86 violates the existing model, thereby exposing the preceding 1 to 7 sequence as an unexplained pattern with high randomness deficiency. Observers are surprised because they are challenged to find a new explanation for a pattern which no longer matches the "ascending sequence" representation.

This example can be formalized as follows using our model of surprise: The black box outputs encoding of strings of the form 1, 2, 3,... n, with a high probability, let's say, of order $1/n^2$. All other sequences of length n (e.g., 1, 2, 3, 4, 5, 6, 7, 86) are outputted with a lower probability of order 2^{-n}. For strings of the first kind, the randomness deficiency is $-\log(1/n^2)$, which is of order $\log n$, minus the shortest description of the string (given the black box distribution function), which is also of order $\log n$ (since a short program could say "print all integers from 1 to n," and n can be encoded in order $\log n$ bits). In the first case, the difference between the probabilistic and computational points of view is small, and so is the surprise. In the second case, the second term stays small, but the first term is $-\log(2^{-n})$, which is linear in n; hence, the difference between the probabilistic and computational points of view is large, and so is the surprise. In this way, any instance of surprise, whether switching from randomness to structure or vice versa, can be presented in terms of patterns being observed without the underlying causal model that would account for them.

Our framework is consistent with explanation-based accounts of surprise (e.g., Foster & Keane, 2015; Maguire, Moser, Maguire, & Keane, 2014; Maguire et al., 2011). Explanation-based accounts propose that surprise is mediated by an estimate of the amount of work needed to integrate an observation with an existing representation, or in other words, the amount of work needed to explain an observation. Results in AIT have shown that the concepts of "explanation" and "data compression" are very closely connected (see Vitányi & Li, 2000): Compressing a sequence of observations is equivalent to explaining those observations, insofar as it permits future observations to be more accurately predicted. The metacognitive estimate of "work" outlined by Foster and Keane (2015) is thus captured by evaluations of randomness deficiency.

It should be noted that our computational approach is not intended to challenge Baldi and Itti's (2010) Bayesian approach. At the limit, both approaches are theoretically equivalent. For example, nonparametric Bayesian approaches have been proposed as a means of explaining how people identify representations that are complex enough to faithfully encode the world, but not so complex as to overfit the data (see Austerweil, Gershman, Tenenbaum, & Griffiths, 2015). What our formalization does differently to the Bayesian approach is simply to provide a convenient means of quantifying surprise in a hypothesis-neutral context. While the Bayesian approach carves up the hypothesis space into a set of discrete alternatives and calculates how beliefs fluctuate between them, our formalization depends only on the specification of a compression scheme that approximates $C(x)$. This allows the surprisingness of an event to be expressed relative to the set of “all other possibilities” without needing to directly represent those alternatives. This approach may prove more useful for modeling open-ended contexts which lack a clear set of salient competing hypotheses. In the case of seeing 66,666 km on a car’s odometer (Desalles, 2006), for example, it seems awkward and unnecessary to express the surprising event as one contrasting with a large number of alternative possibilities (e.g., 66,665 km, 66,664 km, 66,663 km, . . . etc.).

With the formal definition of randomness deficiency in place, we now turn to investigating the role it plays in determining how people respond to real-world situations. In the following section we present an experiment which investigates whether people are sensitive to surprise as we have defined it, and whether they use it to make decisions in practice.

5. Empirical investigation

The following experiment presents a subjectively uncertain scenario that is intuitively amenable to the probabilistic point of view, namely lottery sequences (see Desalles, 2006). While a straight-forward application of probability theory suggests that all lottery sequences are just as likely, their randomness deficiency (i.e., surprisingness) can differ markedly. Rather than needing to specify a set of competing hypotheses, as per Baldi and Itti’s (2010) formulation of surprise, we can express randomness deficiency in terms of data compression.

In the Irish National Lottery, where 6 numbers are drawn from 45, each ordered sequence has a classical probability of $1/C(45, 6) = 1/8.15$ million. According to our theory of surprise, people are sensitive to deviations from randomness and thus should expect the lottery numbers to be Kolmogorov-random (i.e., incompressible), thus requiring $\log_2 8{,}145{,}060 = 23.0$ bits to encode. The more a sequence deviates from a typical random string, the lower the likelihood that it reflects the output of a random source.

Since a universal compressor is an uncomputable ideal (see Li & Vitányi, 2008), we are obliged to create a heuristic compressor which approximates how people experience patterns in lottery sequences. In their investigation into the perceived randomness of binary sequences, Griffiths and Tenenbaum (2004) found that people are sensitive to patterns produced by simple processes such as repetition, symmetry, and duplication (see Bigelow & Piantadosi, 2016; for a large dataset of human-generated number patterns). Based on this research, we developed an encoding scheme which exploits such patterns to reduce description length. This heuristic compressor takes in an ordered sequence of six numbers and computes the six step sizes between them (with the first number counting as the first step). A Huffman encoding scheme is then applied, which relates bit size to step size (a Huffman encoding is one that provides optimal compression under the assumption that all symbols are independent). A breakdown of the structure of the associated Huffman tree is provided in Table 1, with level depth corresponding to the number of bits needed to encode each value.

Table 1
Structure of Huffman encoding scheme

Level Depth	Leaves	No. Branches
1	—	2
2	+1, repeat	2
3	+2, +3	2
4	+4	3
5	+5	5
6	+6	9
7	+7, +8	16
8	+9 up to +40	—

For instance, the sequence 10, 32, 33, 35, 39, 45 is transformed to step sizes of +10, +22, +1, +2, +4, +6, which is then encoded using 8 + 8 + 2 + 3 + 4 + 6 = 31 bits. Using this scheme, an analysis of 6 years of bi-weekly Irish National Lottery draws revealed a mean compressed length of 30.9 bits (*SD* = 3.6), with a mode of 31 bits. The most randomness-deficient of the 624 sequences was 2, 4, 32, 34, 36, 37 (description length of 20 bits), while the most random was 9, 20, 26, 27, 34, 45 (description length of 39 bits). The theoretical minimum description length of our system is 12 (e.g., 1, 2, 3, 4, 5, 6), while the theoretical maximum is 43 (e.g., 7, 13, 20, 29, 36, 45). This contrasts with the 23.0 bits needed to perfectly encode an ordered random sequence of six numbers between 1 and 45.

Although our compressor does not capture all computable patterns, it delivers compression for randomness-deficient outputs (i.e., it compresses below 23.0 bits for certain non-typical random sequences) and can therefore be used to evaluate the hypothesis that people use randomness deficiency to adjust beliefs regarding causal origin. It should be noted that our choice of encoding is just one possible compression scheme among many that could be adopted, all of which would presumably yield similar results.

5.1. *Experiment*

Two quickpick (i.e., randomly selected) lottery tickets were purchased for the subsequent week’s Irish National Lottery draw, each with six ordered numbers ranging from 1 to 45. The number on the first ticket had a compressed description length of 31 bits (2, 8, 11, 19, 23, 38), while the second had a length of 30 bits (6, 13, 17, 20, 43, 44). Based on our analysis of 5 years of winning tickets, these lengths were close to the mean/modal description lengths of Irish lottery ticket sequences. Each of the number sequences from these two tickets were, respectively, combined with four other number sequences that had been specifically constructed to have different compression lengths, and thus different levels of surprisingness. The aim of the experiment was to identify whether participants would rely on randomness deficiency to identify the original lottery sequences in a hypothesis-neutral context.

5.1.1. Participants

One hundred thirty undergraduate students from Maynooth University participated voluntarily in this study.

5.1.2. Procedure

Participants were informed that a pair of valid lottery tickets had been purchased for that week's draw, which would be displayed at the end of the experiment. They were presented with the two sets of five number sequences (i.e., one of the lottery number sequences mixed in with the four constructed sequences). Their goal was to identify the lottery sequence from within the set of five candidate sequences. No mention was made of how the other four constructed sequences had been obtained.

Each quickpick sequence was presented on a screen along with four other sequences generated randomly by a computer algorithm. This program kept iterating through random sequences until finding one matching the required bit size. The four distractor sequences met the constraints of having compressed bit-sizes of 15–18 bits (e.g., 2, 4, 5, 6, 10, 12), 19–22 bits (e.g., 14, 15, 18, 19, 20, 22), 23–26 bits (e.g., 10, 11, 22, 23, 24, 27), and 27–29 bits (e.g., 7, 11, 15, 16, 18, 44), respectively. The ordering of the five sequences on the screen was randomized.

To avoid influencing participants into thinking in a particular way about the task, we deliberately avoided using the words "surprise" or "probability." Participants were instead asked to rank each set of five sequences according to perceived likelihood of being the quickpick sequence, from highest likelihood to lowest likelihood.

Unfortunately for the experimenters, the lottery tickets did not turn out to be winning ones.

5.2. Results

An individual applying classical probability would view all sequences as equally likely and would thus only have a 20% chance of correctly identifying one quickpick sequence mixed with four distractor sequences. However, 64% of participants correctly identified the numbers on the first ticket, and 66% on the second ticket (i.e., ranked these sequences in first place out of the five possibilities). The first ticket had a compressed description length of 31 bits, while the second had a length of 30 bits.

Fig. 1 shows the mean compressed bit size for sequences ranked from first to fifth place across the two presentations. The overall correlation between ranking and compressed description length was 0.965, $p < .001$.

For judgments involving the first ticket, the mean compressed description lengths of sequences ranked in places from first to fifth were 28.2, 25.6, 24.1, 21.8, and 18.4 bits, respectively. A Friedman test revealed a significant difference in the bit sizes of perceived likelihood ranks, $\chi^2(4) = 162.8$, $p < .01$. Post hoc analysis using Wilcoxon signed-rank tests with Bonferroni corrections found significant differences between all comparisons ($p < .01$).

For judgments involving the second ticket, the mean compressed description lengths were 28.4, 25.7, 23.1, 21.1, and 19.4 bits, respectively. These results again revealed a significant

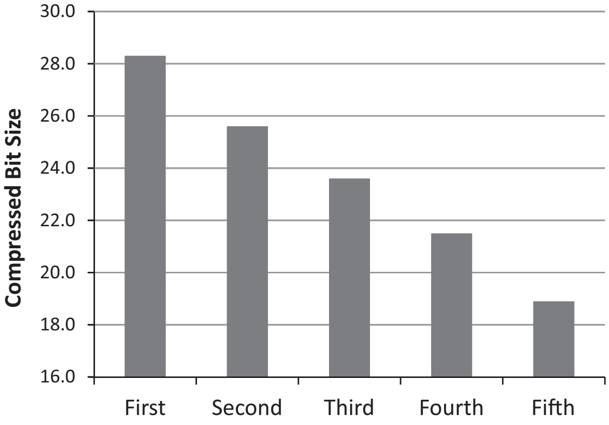

Fig. 1. Mean compressed bit size according to rankings of likelihood.

difference in perceived likelihood ranks, $\chi^2(4) = 209.1$, $p < .01$. Again, all comparisons were shown to differ in post-hoc analysis using Bonferroni-adjusted Wilcoxon signed-rank tests, with larger bit size sequences being ranked more likely to be the lottery sequence.

5.3. Discussion

Our results demonstrate that, not only are people sensitive to randomness deficiency, they rely on it to enhance their judgement accuracy. While probability theory assumes that all lottery sequences are equally likely, people realize that there is an element of uncertainty involved in how those sequences have been generated. They understand that the greater the randomness deficiency of a lottery sequence, the greater the likelihood that it was produced by an alternative non-random mechanism. These results underscore the importance of surprise in reasoning and decision making.

Although the Bayesian approach and the computational approach are theoretically equivalent at the limit, our model may offer closer insight into the experience of "surprise detection" (as opposed to "surprise calculation"). As argued by Foster and Keane (2015), the initial surprise response reflects a metacognitive estimate of the work involved in explaining an abnormal event. This initial estimate is more likely to involve a basic hypothesis-neutral appraisal of events, as described by our computational model of surprise, as opposed to a more fine-grained analysis of how beliefs are shifted among a specific set of hypotheses, which Baldi and Itti's (2010) quantification entails. Our model demonstrates how an event can be perceived as anomalous even in the absence of any salient hypotheses (e.g., repeated numbers in the Bulgarian lottery). Because randomness deficiency acts as a context-neutral indicator of suboptimal representation, it provides a valuable heuristic for homing in on potentially informative stimuli.

Converging evidence points to a connection between randomness deficiency and learning. Infants, for example, display a sensitivity to coincidence, selectively exploring objects that produce anomalous data (Gopnik & Schulz, 2004; Xu & Kushnir, 2013), as well as relying on suspicious coincidence to infer the correct meaning of words (Jenkins, Samuelson, Smith, & Spencer, 2014). The detection of surprising events provokes animals to learn faster (Courville, Daw, & Touretzky, 2006), and it has also proved key to causal discovery and rational inference in the sciences (Xu & Kushnir, 2013). People are fascinated by the experience of unexpected patterns (Itti & Baldi, 2006) and can accurately assess the level of statistical support they provide for an underlying causal structure, thus equipping them with a fundamental skill for concept learning and theory formation (Griffiths & Tenenbaum, 2007). Taken to unproductive extremes, however, this responsiveness to anomalous patterns produces a condition known as “apophenia” (see Fyfe, Williams, Mason, & Pickup, 2008).

In sum, people’s attention naturally gravitates toward subject matter which offers the potential for the identification of unexpected patterns, setting up a close link between surprise, randomness deficiency, curiosity, and interestingness. Schmidhuber (2009), for instance, argues that the experience of randomness deficiency, with subsequent resolution through representational updating (i.e., data compression), is what makes subjects interesting, films entertaining, and jokes funny. Both scientists and artists actively select experiments in search of simple innovative laws which would compress the observation history, thereby leading to enhanced understanding. Schmidhuber (2009) argues that the creativity of painters, dancers, musicians, pure mathematicians, and physicists can all be viewed as the by-product of a human drive toward enhanced compression progress.

6. Conclusion

No matter how hard we try, it is never possible to eliminate uncertainty. For highly specialized and precisely engineered situations, such as those involving games of chance, representations are so reliable that the possibility of surprise can be effectively ignored. In noisy real-world environments, however, surprise is a common experience, and failing to identify it can have detrimental consequences.

In this article, we have provided a general model of surprise. Rather than expressing it in terms of shifting hypotheses (e.g., Baldi & Itti, 2010), we have defined surprise in terms of universal pattern detection. The observation of randomness deficiency is a good heuristic for a representation that is missing something, since compressible patterns rarely occur by chance and are much more likely to be driven by some underlying structure which is not being correctly modeled. For this reason, it makes sense for people to be constantly alert to the appearance of patterns in unusual contexts.

References

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford* *handbook of computational and mathematical psychology* (pp. 187–208). Oxford, UK: Oxford University Press.

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, *23*, 649–666.

Bigelow, E., & Piantadosi, S. (2016). A large dataset of generalization patterns in the number game. *Journal of Open Psychology Data*, *4*(1), 4.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566.

Chater, N., & Brown, G. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science*, *32*, 36–67.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 811–823.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.

Courville, A. C., Daw, N. D., – Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, *10*(7), 294–300.

Desalles, J. L. (2006). A structural model of intuitive probability. In D. Fum, F. Del Missier, & A. Stocco (Eds.), *Proceedings of the seventh international conference on cognitive modeling* (pp. 86–91). Trieste, IT: Edizioni Goliardiche.

Dessalles, J. L. (2008). Coincidences and the encounter problem: A formal account. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the thirtieth annual conference of the Cognitive Science Society* (pp. 2134–2139). Austin, TX: Cognitive Science Society.

Foster, M. I., & Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, *81*, 74–116.

Friston, K., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, *3*, 151.

Fyfe, S., Williams, C., Mason, O. J., & Pickup, G. J. (2008). Apophenia, theory of mind and schizotypy: Perceiving meaning and intentionality in randomness. *Cortex*, *44*(10), 1316–1325.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, *8*, 371–377.

Griffiths, T., & Tenenbaum, J. (2004). From algorithmic to subjective randomness. *Advances in Neural Information Processing Systems*, *16*, 953–960.

Griffiths, T., & Tenenbaum, J. (2007). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems*, *19*, 547–554.

Jenkins, G. W., Samuelson, L. K., Smith, J. R., & Spencer, J. P. (2014). Non-bayesian noun generalization in 3-to 5-year-old children: Probing the role of prior knowledge in the suspicious coincidence effect. *Cognitive Science*, *39*, 268–306.

Li, M., & Vitányi, P. (2008). *An introduction to Kolmogorov complexity and its applications. Texts in Computer Science (Vol. 9)*. New York: Springer.

Loewenstein, J., & Heath, C. (2009). The repetition-break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive Science*, *33*(1), 1–19.

Maguire, P., & Maguire, R. (2009). Investigating the difference between surprise and probability judgments. In N. A. Taatgen & H. V. Rijn (Eds.), *Proceedings of the Thirty-First Annual Meeting of the Cognitive Science Society* (pp. 2539–2564). Austin, TX: Cognitive Science Society.

Maguire, R., Maguire, P., & Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(1), 176–186.

Maguire, P., Moser, P., Maguire, R., & Keane, M. (2014). A computational theory of subjective probability. *arXiv preprint arXiv:1405.6142*.

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*(3), 251–274.

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: Kaufmann.

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134.

Schauerte, B., & Stiefelhagen, R. (2013). "wow!" bayesian surprise for salient acoustic event detection In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 6402–6406). IEEE.

Schmidhuber, J. (2009). Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of the Society of Instrument and Control Engineers*, *48*(1), 21–32.

Sloman, S. A., & Lagnado, D. (2005). The problem of induction. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking & reasoning* (pp. 95–116). New York: Cambridge University Press.

Susskind, J. M., Lee, D. H., Cusi, A., Feiman, R., Grabski, W., & Anderson, A. K. (2008). Expressing fear enhances sensory acquisition. *Nature Neuroscience*, *11*(7), 843–850.

Teigen, K. H., & Keren, G. (2003). Surprises: Low probabilities or high contrasts? *Cognition*, *87*(2), 55–71.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., & Nelson, C. (2009). The nimstim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*(3), 242–249.

Vitányi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *46*, 446–464.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*(1), 28–32.

Zhao, J., Hahn, U., & Osherson, D. (2014). Perception and identification of random events. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1358.