Unexpectedness and Bayes' Rule

Giovanni Sileno[1] and Jean-Louis Dessalles[2]

[1] University of Amsterdam, Amsterdam, The Netherlands
[2] Télécom Paris, Paris-Saclay University, France
g.sileno@uva.nl, dessalles@telecom-paris.fr

Abstract. A great number of methods and of accounts of rationality consider at their foundations some form of Bayesian inference. Yet, Bayes' rule, because it relies upon probability theory, requires specific axioms to hold (e.g. a measurable space of events). This short document hypothesizes that Bayes' rule can be seen as a specific instance of a more general inferential template, that can be expressed also in terms of algorithmic complexities, namely through the measure of unexpectedness proposed by Simplicity Theory.

Keywords: Bayes' rule· Unexpectedness · Algorithmic Complexity · Simplicity Theory · Computational Cognitive Model

1 Introduction

Since its introduction in philosophy and mathematics to analyse chances in games, probability theory has grown to be one of the most important ingredients of formal accounts of how rational agents (artificial or natural) should reason in conditions of uncertainty. Central to this enterprise is the famous Bayes' rule, at the base of Bayesian models (a family including Bayesian networks), Bayesian inference, *maximum a posteriori* (MAP) estimation in statistics, and core component of various machine learning methods (e.g. *variational autoencoders* [13]). Besides being part of the common toolkit to support or reproduce human decision-making (e.g. for medical diagnosis [18], for evidential reasoning in legal cases [10], see also [9]), Bayesian models have been applied in cognitive sciences to topics as diverse as animal learning [3], visual perception [20], motor control [14], language processing [2], and forms of social cognition [1]. Such a success can be explained by the clarity of the theoretical framework, and the undoubted practical value it has proven in several application domains. However, reasons exist for which Bayesian inference may be neither a descriptive, nor a prescriptive model of human reasoning.

As a formal framework, probability theory relies on a series of axioms to hold (e.g. a measurable space of events), which enables a closure provided with interesting mathematical properties, but which is not necessarily representative of the way in which humans mentally form or process events. Given any description of the world we may always find a description which differs in some aspect from the previous one, adding any detail. As a modeling framework, the limitations

of standard theory of probability to capture human reasoning is proven by the existence of several cognitive patterns (often named biases or fallacies) which do not follow what is predicted by the formal theory, see e.g. [12, 19]. The core limitation motivating the present contribution lies however in the *mismatch* between what humans see as *informative* and the definition of information given by Shannon, that triggered in the '90s the introduction of Simplicity Theory (ST) [4]. The present paper introduces a novel hypothesis concerning the theoretical bases which makes this cognitive model functional.

1.1 Simplicity Theory

Simplicity Theory (ST) is a computational model of cognition found to predict diverse human phenomena related to relevance (unexpectedness [6], narrative interest [8], coincidences [7], near-miss experiences [5], emotional interest [15], responsibility [17]), used also for experiments in artificial creativity [16]. Core contributions of ST are: (a) a non-extensional theory of *subjective probability*, centered around the notion of unexpectedness; (b) a model of emotional intensity predicting emotional amplification in occurrence of unexpected phenomena. For our aims here, we will focus only on the (a) part. Formally, ST builds upon results obtained in *algorithmic information theory* (AIT), (see e.g. [11]).

Kolmogorov Complexity In AIT, the *complexity* of a string x is the minimal length of a program that, given a certain optional input parameter y, produces x as an output:

$$K_\phi(x|y) = \min_p \{|p| : p(y) = x\}$$

The length of the minimal program depends on the operators and symbols available to the computing machine ϕ.[3] If specified on universal Turing machines, this measure is generally incomputable, and it is defined always up to a constant. If the machine is resource-bounded, complexity is computable; the bounded version will be here denoted as C. This definition of complexity can be mapped to any domain, as long as one defines what are the *symbols* and the computations that are performed on these symbols; under certain conditions, the search for the minimal program can be mapped to min-path or functionally similar algorithms.

Unexpectedness ST's measure of *unexpectedness* (U) is defined as the divergence between two resource-bounded Kolmogorov complexities: the *causal* (also *world*, or *generation*) *complexity* C_W and the *description complexity* C_D:

$$U(s) = C_W(s) - C_D(s)$$

where s is a *situation* parameter. In various experiments, this measure has proven to predict shortcomings of standard theory of information observable in everyday

[3] K is an (algorithmic) *informational complexity*: it captures how much information is needed for constructing the object, but not how much time or space is needed (it is distinct from the algorithmic/time complexity used to study tractability).

life. Examples include e.g. remarkable lottery draws (e.g. 11111 is more unexpected than 64178, even if the lottery is fair), coincidence effects (e.g. meeting by chance a friend in a foreign city is more unexpected than meeting any unknown person equally improbable), deterministic yet unexpected events (e.g. a lunar eclipse), and many others [4, 7, 5, 6]. Representing diagrammatically the domains of the two complexities underlying unexpectedness, we have:

$$\overbrace{\text{world} \rightarrow \text{situation}}^{C_W} \qquad \overbrace{\text{situation} \leftarrow \text{mind}}^{C_D}$$

2 Unexpectedness and Bayes' Rule

Our aim here is to provide further arguments in support to non-probabilistic computational models of cognition, in particular focusing on the following:

Conjecture. *Bayes' rule is a specific implementation of a more general inferential template, captured by ST's definition of unexpectedness.*

To construct this claim, we start from the definition of *conditional probability*:

$$p(O \cap M) = p(M|O) \cdot p(O) = p(M) \cdot p(O|M)$$

where O denotes an observation, and M a model (both elements from the same measurable space). Bayes' formula is:

$$p(M|O) = \frac{p(M \cap O)}{p(O)} = \frac{p(O|M) \cdot p(M)}{p(O)}$$

The formula is often expressed using informal terms:

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

Now, empirical observations [7] suggest that U can be put in correspondence to posterior probability, i.e.

$$\text{posterior} = 2^{-U}$$

This entails that when $U \approx 0$ (posterior ≈ 1), the situation confirms the agent's model of the world (it is *plausible*) and therefore it is not informative. (Note that to maintain a correspondence with probabilities, U needs also to be superior or at least equal to 0.) However, we tacitly overlooked a detail. Unexpectedness has only a parameter s, whereas posterior probability refers to O and M. Intuitively, s corresponds to O and not to M: as an observation concerns the situation in focus, possibly perceived as unexpected. But then, where can we find M? In order to understand this absence, let us reconsider Bayes' formula. Inverting the terms of the equation, and using the logarithm, we can form a mapping to unexpectedness, i.e.:

$$\overbrace{\log \frac{1}{p(M|O)}}^{U(s)} = \log \frac{p(O)}{p(O|M) \cdot p(M)} = \overbrace{\log \frac{1}{p(O|M)} + \log \frac{1}{p(M)}}^{C_W(s)} - \overbrace{\log \frac{1}{p(O)}}^{C_D(s)}$$

Causal complexity Let us start from $C_W(s)$, the *causal complexity*, i.e. the length in bits of the shortest path that, according to the agent's world model, generates the situation s. If s is a phenomenon, an *event* probabilistically captured by O, s can be seen as the manifestation of some pre-existing causal mechanisms c, that probabilistically is captured with M. Then, in order to generate s (e.g. the symptoms of a disease), the world has first to generate its cause c (e.g. the disease), expressing the application of a *chain rule*:

$$C_W(s) \rightsquigarrow C_W(c * s) = C_W(s||c) + C_W(c)$$

where $C_W(s||c)$ is the complexity of generating s from a state of the world in which c is the case, and $c*s$ is the sequential chaining of c and s ('$||$' and '$*$' add temporal contraints that '$|$' and '$\cap$' in probability formulas do not have). From the definition of Kolmogorov complexity, the mapping is an equality if and only if the shortest path to s passes from c, i.e. if c is the *best explanation* of s:

$$C_W(s) = \min_c C_W(c * s) = \min_c [C_W(s||c) + C_W(c)]$$

Therefore the unexpectedness formula can be seen as abstracting the *causally explanatory* factor c, with the implicit assumption that the best one is automatically selected in the computation of complexity.

Description complexity Additionally, ST specifies C_D, the description complexity, as the length in bits of the shortest determination of the object s. Such shortest determination may consist e.g. in specifying the address where to retrieve it from memory. Note that from a computational point of view, U could be negative, namely when the description of s is more complex than its generation; we are in this case in front of *inappropriate* descriptions, as they are adding irrelevant information for their function.

In the terms suggested by Bayes' formula, C_D corresponds to the probability of having *observed* a certain situation. The link between descriptive complexity and probability can be then established through *optimal encoding* in Shannon's terms, where probability is assessed through frequency ($\log \frac{1}{p(O)}$). However, this approach does not take into account possible mental compositional effects (e.g. Gestalt-like phenomena), nor events that never occurred before. Complexity is a more generally applicable measure.

Comparison with Bayes' rule The previous observations allows us to claim that Bayes' rule is a specific instantiation of ST's Unexpectedness that: (a) makes a candidate "cause" explicit and does not select automatically the best candidate; (b) takes a frequentist-like approach for encoding observables. More formally:

$$U(s) = \min_c \overbrace{[C_W(c * s) - C_D(s)]}^{\text{posterior}} = \min_c [\overbrace{C_W(s||c)}^{\text{likelihood}} + \overbrace{C_W(c)}^{\text{prior}} - \overbrace{C_D(s)}^{\text{evidence}}] \quad (1)$$

Note that this formula relies on the explicit assumption that c precedes s (as indicated by the symbols $*$ and $||$). This restriction is absent from Bayes' rule, in which the model M and the observation O can exchange roles; their causal dependence does not lie in the rule, but solely in the eye of the modellers.

3 All Prior is Posterior of Some Other Prior?

By accepting the previous mapping, we find ourselves in front of a dilemma. Probability functions are functions of the same type, independently on whether they are prior or posterior, whereas for instance complexity of description (that maps to evidence in Bayes' terms) and unexpectedness (to posterior) are not.

Let us consider an additional prior in Bayes' formula (a sort of *contextual* prior), denoted with E (standing for 'environmental context'):

$$p(M|O,E) = \frac{p(M \cap O|E)}{p(O|E)} = \frac{p(O|M,E) \cdot p(M|E)}{p(O|E)}$$

Following probability theory, an equivalent form for computing the posterior would be considering the composite event $O \cap E$:

$$p(M|O,E) = \frac{p(M \cap O \cap E)}{p(O \cap E)}$$

These two formulations, rewritten in terms of complexities, are not equivalent. First, a sequential chaining of situations (e.g. $e * s$, using e for environmental situation) is not the same as an unordered conjunction of random events ($O \cap E$). The environmental situation e (the context) has to be generated *before* the situation c (the cause), which in turns occurs *before* the target situation s (the effect). Accepting these temporal constraints, the second expression can be mapped to:

$$\frac{p(M \cap O \cap E)}{p(O \cap E)} \qquad \rightsquigarrow \qquad C_W(e * c * s) - C_D(e * s)$$

As before, c can disappear when it is assumed to be part of the "best available" course of events to produce s from e. By doing this, the measure becomes equivalent to the unexpectedness of the chaining of s after e:

$$C_W(e * s) - C_D(e * s) = U(e * s)$$

In contrast, the conditional version expression of the probability ratio refers to a distinct computation:

$$\frac{p(M \cap O|E)}{p(O|E)} \qquad \rightsquigarrow \qquad C_W(c * s||e) - C_D(s|e)$$

where $C_D(s|e)$ is the complexity of describing s when e is given as input. Taking c out, the formula suggests introducing the notion of conditional or hypothetical unexpetedness:

$$C_W(s||e) - C_D(s|e) \equiv U(s||e)$$

This definition is in line with the fact that the descriptive (e.g. conceptual) remoteness of s from e, expressed by the term $C_D(s|e)$, discounts the unlikelihood of their causal connection (related to the improbability of O given E), making it more plausible (less unexpected).

As we can see, ST makes a distinction between two versions of the probabilistic conditional $p(M|O,E)$ that probability calculus conflates. This leads us to considering notions such as framing and relevance.

3.1 Informational Principle of Framing

The difference between the two formulations of the posterior $p(M|O,E)$, when mapped to complexities, can be computed as:

$$U(e*s) - U(s||e) = C_W(e*s) - C_D(e*s) - C_W(s||e) + C_D(s|e)$$

The use of chaining $(*)$ within C_W shares similar form than the chain rule in probability:

$$C_W(e*s) = C_W(e) + C_W(s||e)$$

that is, in order to generate e and then s, the world needs first to generate e (from the current configuration), and then to generate s in a configuration in which e has been generated. Instead, the chain rule for the description complexity C_D depends on the description machine and provides us only an upper bound:

$$C_D(e*s) \leq C_D(e) + C_D(s|e)$$

This is because we do not have the temporal constraints, and the minimal path for describing e and s together may turn out to be simpler than a constrained path in which one term is fully determined before the other. Applying the two chain rules we have:

$$U(e*s) - U(s||e) \geq C_W(e) - C_D(e) = U(e)$$

Thus, a necessary condition for which the two formulations may be equivalent is that $U(e) = 0$, i.e. when the contextual prior is *not unexpected*. This constraint implicitly brings forward an *informational principle of framing*: all contextual situations e which are not unexpected (shared facts, defaults, but also improbable but descriptively complex situations) provide grounds to be neglected. The remaining situations provide the "relevant" context for the situation in focus.

4 Likelihood and Prediction

Suppose we want to predict *ex-ante* a certain outcome, given certain circumstances. In probabilistic terms, the relevant measure for prediction is the likelihood $p(O|M)$. The conjecture expressed above suggests that likelihood matches ST's notion of conditional causal complexity: $C_W(s||c)$ (where M and c play the role of contextual priors).

However, ST's framework also suggests that humans have limited access to C_W. When there are n options playing symmetrical roles, it seems there is no difficulty to measure $C_W = \log_2(n)$. Otherwise, people tend to imagine a situation in which s occurred in order to measure its likelihood. To do so, s needs to be adequately framed, and therefore there needs to be some calculation of C_D, so in this case there cannot be C_W without C_D. This implies that the assessment of the likelihood probability $p(O|M)$ is indirect in ST. Let's call $C_W^U(s||c)$ the causal complexity *derived* from unexpectedness:

$$C_W^U(s||c) = U(s||c) + C_D(s|c)$$

The formula captures the fact that the conceptual remoteness of s from c this time *adds* to the unexpectedness (implausibility) of observing their connection, making this connection less likely (more improbable).

Examples Consider the estimation of the likelihood that the wall changes colour (s) if I close the door (c). The wall is part of perceptions, therefore its determination is immediate ($C_D \approx 0$), so $U \approx C_W \gg 0$ (because I never experienced something similar). It would then be highly unexpected if it occurred. The likelihood would also be very low, as the derived causal complexity is very high:

$$C_W^U = U + C_D \approx U + 0 \gg 0$$

Now suppose that someone tells me that there is a special light projector commanded by the door state. C_W would drop, as well as the posterior U, and in turn the derived likelihood C_W^U.

Consider instead the likelihood that, when I close the door, a certain stone somewhere in the world moves. The complexity for determining that specific stone is high, i.e. $C_D \gg 0$. The causal complexity of seeing that specific stone moving is also very high $C_W \gg 0$, therefore we have $U \approx 0$: it is plausible that some essentially random stone may move at the moment I open the door. However, the resulting likelihood is still very high, because:

$$C_W^U = U + C_D \approx 0 + C_D \gg 0$$

If we had $C_D = 0$, the likelihood would be just the same as the posterior. If, in the stone example, the portion of the world we look at includes the stone (e.g. in front of us), C_D is reduced (up to ≈ 0), increasing U, but maintaining the same value of C_W^U. A similar consideration applies if we repeat the experiment twice with the same remote stone (we do not need to describe it again).

5 Posterior and Post-diction

Suppose that we want to retrodict or abduce certain circumstances given a certain outcome, or *ex-post*. From a probabilistic perspective, this amounts to computing the posterior $p(M|O)$. Following the conjecture expressed above, this corresponds to computing $U(s)$, if the cause c lies in the generative path bringing to s. But what if c is not part of that path? On some occasions, one may want to compute the complexity of an alternative path in which c plays a role. Looking back at the conjecture expressed in (1), this can be captured via a *causally constrained* unexpectedness $U_c(s)$, where c is the constraining cause:

$$U_c(s) = C_W(c * s) - C_D(s) \qquad U(s) = \min_d U_d(s)$$

Note that the cause does not play an explicit role in the computation of the description complexity. Then we have:

$$U_c(s) - U(s) = \min_d [C_W(s||c) - C_W(s||d) + C_W(c) - C_W(d)] \geq 0$$

However, when the cause is described explicitly, the observer has to consider the full sequence, and this corresponds to computing $U(c*s)$, with $U(c*s) \leq U_c(s)$.

Prosecutor's Fallacy Suppose that, following forensic studies, the likelihood $p(O|M)$ (e.g. the probability that a certain DNA evidence appears if the defendant is guilty) is deemed very high, i.e. $p(O|M) \approx 1$. The *prosecutor's fallacy* [19] occurs when the posterior $p(M|O)$ (the probability that the defendant is guilty given that there is DNA evidence) is also concluded to be comparatively high:

$$p(O|M) \approx 1 \rightsquigarrow p(M|O) \approx 1 \qquad \text{[Prosecutor's fallacy]}$$

This is a fallacy, because the correct criterion for applying this reasoning pattern would be that the priors compensate each other, i.e. $p(M) \approx p(O)$.

Now, let us look at the same scenario in terms of complexity. For the conjecture, the posterior $p(M|O)$ maps to $U(s)$, if c lies in the causal path; to $U_c(s)$ in the general case. The likelihood $p(O|M)$, on the other hand, maps to $C_W(s||c)$. Let us retake the definition of $U_c(s)$:

$$U_c(s) = C_W(c * s) - C_D(s) = C_W(s||c) + C_W(c) - C_D(s)$$

Knowing that $C_W(s||c) \approx 0$ (the causal connection is deemed strong), $U_c(s)$ can be zero only if $C_W(c) \approx C_D(s)$. If the cause is not unexpected — it is deemed plausible from the prosecutor standpoint (e.g. an adequate explanation can be found of how the defendant was there), we have $U(c) = C_W(c) - C_D(c) \approx 0$. In other words, the prosecutor's fallacy emerges if the complexity of description of outcome (e.g. evidence) and cause (e.g. being guilty) are comparable, i.e. $C_D(c) \approx C_D(s)$, which seems a sound hypothesis considering the usually limited list of suspects in the mind of the prosecutor.

6 Conclusion

Our conjecture that Bayes' rule is a specific form of a more general inferential template provides further arguments in support to non-probabilistic computational models of cognition. A complexity-based account of the posterior allows distinguishing between relevant and irrelevant contextual elements, while the probabilistic account treats them equally. Acknowledging that measures of bounded complexity are computable, the question becomes then *how* the underlying machines should be defined, for developing computational agents, or with the purpose of modeling human cognition. Yet, the abstraction level of algorithmic information complexity is already relevant to draw conclusions about expected outcomes, even without looking at internal workings. This opens the possibility of novel insights, as we have shown here for instance with the analysis of the prosecutor's fallacy.

References

1. Baker, C.L., Tenenbaum, J.B.: Modeling human plan recognition using bayesian theory of mind. In: Sukthankar, G., Geib, C., Bui, H.H., Pynadath, D.V., Goldman, R.P. (eds.) Plan, Activity, and Intent Recognition, pp. 177–204. Morgan Kaufmann, Boston (2014)
2. Chater, N., Manning, C.D.: Probabilistic models of language processing and acquisition. Trends in Cognitive Sciences **10**(7), 335–344 (2006)
3. Courville, A.C., Daw, N.D., Touretzky, D.S.: Bayesian theories of conditioning in a changing world. Trends in Cognitive Sciences **10**(7), 294–300 (2006)
4. Dessalles, J.L.: La pertinence et ses origines cognitives. Hermes-Science (2008)
5. Dessalles, J.l.: Simplicity Effects in the Experience of Near-Miss. COGSCI 2011, Annual Meeting of the Cognitive Science Society pp. 408–413 (2011)
6. Dessalles, J.L.: Algorithmic simplicity and relevance. Algorithmic probability and friends **7070 LNAI**, 119–130 (2013)
7. Dessalles, J.L.J.L.: Coincidences and the encounter problem: A formal account. Cognitive Science pp. 2134–2139 (2011)
8. Dimulescu, A., Dessalles, J.l.: Understanding Narrative Interest : Some Evidence on the Role of Unexpectedness. In: Proceedings of the 31st Annual Conference of the Cognitive Science Society. pp. 1734–1739 (2009)
9. Fenton, N., Neil, M.: Risk assessment and decision analysis with Bayesian networks. Crc Press (2018)
10. Fenton, N., Neil, M., Lagnado, D.A.: A general structure for legal arguments about evidence using bayesian networks. Cognitive science **37**(1), 61–102 (2013)
11. Grünwald, P.D., Vitányi, P.M., et al.: Algorithmic information theory. Handbook of the Philosophy of Information pp. 281–320 (2008)
12. Kahneman, D., Slovic, S., Slovic, P., Tversky, A., Press, C.U.: Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press (1982)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. Proceedings of 2nd International Conference on Learning Representations (ICLR2014) (2014)
14. Körding, K.P., Wolpert, D.M.: Bayesian decision theory in sensorimotor control. Trends in Cognitive Sciences **10**(7), 319–326 (2006)
15. Saillenfest, A., Dessalles, J.L.: Role of Kolmogorov Complexity on Interest in Moral Dilemma Stories. In: Proc. of the 34th Annual Conf. of the Cognitive Science Society. pp. 947–952 (2012)
16. Saillenfest, A., Dessalles, J.L., Auber, O.: Role of simplicity in creative behaviour: The case of the poietic generator. Proceedings of the 7th International Conference on Computational Creativity, ICCC 2016 pp. 33–40 (2016)
17. Sileno, G., Saillenfest, A., Dessalles, J.L.: A Computational Model of Moral and Legal Responsibility via Simplicity Theory. Proc. of the 30th Int. Conf. on Legal Knowledge and Information Systems (JURIX 2017) **FAIA 302**, 171 – 176 (2017)
18. Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., Cowell, R.G.: Bayesian analysis in expert systems. Statistical science pp. 219–247 (1993)
19. Thompson, W.C., Schumann, E.L.: Interpretation of statistical evidence in criminal trials. Law and Human Behavior **11**(3), 167–187 (1987)
20. Yuille, A., Kersten, D.: Vision as bayesian inference: analysis by synthesis? Trends in cognitive sciences **10**(7), 301–308 (2006)