https://doi.org/10.1038/s44159-026-00570-1

The explicit-Bayes hypothesis for cognition

Kai Xue & Dobromir Rahnev

Check for updates

It is often asserted that human cognition is Bayesian, but that broad claim is difficult to test in a falsifiable way. We suggest that researchers specifically assess a narrower, falsifiable hypothesis: that cognition explicitly uses Bayes' rule at the algorithmic level.

A widely adopted approach to understanding human cognition is to consider it to be Bayesian inference[1]. Bayesian inference is the process of computing the probability of a hypothesis given observed data, specifically by combining prior probabilities with the likelihood of the data under that hypothesis (using a formula known as Bayes' rule). This approach connects cognition to a well-defined standard of rationality and offers a unified language for describing cognition across diverse domains. However, the idea that cognition is Bayesian inference has also been criticized for being overly flexible and unfalsifiable[1]. The debate has stalled in part due to persistent confusion about the level at which claims about Bayesian inference are intended to apply.

Levels of analysis

Marr's three levels of analysis can be used to clarify what 'cognition is Bayesian' can mean. Marr proposed that any problem can be considered at three levels of analysis: computational, algorithmic, or implementational[2]. The computational level describes the problems the mind needs to solve, the algorithmic level specifies how a computation is carried out and the implementational level defines the neural mechanisms that realize the computation. The three levels are related but separate: specifying the goal of a computation does not by itself determine the algorithm that carries it out, and specifying an algorithm does not by itself determine how it is physically implemented. For instance, the computational level defines the goal of recognizing a spoken word, the algorithmic level defines the process by which acoustic features are weighted and combined, and the implementational level defines the neural circuits in auditory cortex that carry out that process.

The view that cognition is Bayesian inference is often intended as a statement about the computational level[3]. In other words, it is best interpreted as a framework that explains why cognition is the way it is rather than as a falsifiable hypothesis about the underlying algorithms[3]. However, this distinction between computations and algorithms has proven hard to maintain in practice. When a Bayesian model shows a good fit to behavioural data, it is tempting to interpret that result as evidence for explicit use of Bayes' rule at the algorithmic level. This tendency is exacerbated by the fact that most models in cognitive science are intended to be interpreted at the algorithmic level, thus making it more likely that a Bayesian model is interpreted at that level too.

To explicitly separate the computational and algorithmic levels, we propose considering Bayesian inference separately at each level. Thus, in addition to considering Bayesian inference at the computational level, a separate but related objective for the field should be to test whether the underlying algorithms use Bayes' rule.

Bayes at the algorithmic level

A strength of assessing Bayesian cognition at the algorithmic level is that much of cognitive science already focuses on this level. Focusing on the algorithmic level reduces confusion about whether any particular claim about Bayesian inference is meant to apply only to the computational level or also to the algorithmic level. For example, 'fast and frugal' heuristics (such as only using the most informative factor and ignoring the rest) are non-Bayesian at the algorithmic level[4] even though they can be conceptualized as Bayesian inference under extreme priors at the computational level[5].

We refer to the hypothesis that cognitive algorithms explicitly use Bayes' rule as the 'explicit-Bayes hypothesis'. It is possible for an explicit-Bayes model to incorporate priors that do not reflect the environment, assume wrong rewards, or generally mis-specify any quantity used in applying Bayes' rule. The criterion for what makes a model explicit-Bayes is simply that it includes the explicit application of Bayes' rule. Although this hypothesis is only one of many Bayes-relevant hypotheses that one could formulate at the algorithmic level, it has multiple advantages.

Most importantly, the explicit-Bayes hypothesis is both unambiguous and falsifiable. Alternative criteria for distinguishing Bayesian and non-Bayesian algorithms, such as representation of uncertainty or computational rationality, do not draw a clear empirical boundary between Bayesian and non-Bayesian models — many cognitive models can be seen as incorporating uncertainty or being rational in some sense.

Under the criterion adopted here, certain models that rely on sampling, amortized inference, or predictive coding would count as non-explicit-Bayes models even though they were specifically designed as Bayesian models at the computational level. This refinement clarifies the nature of these algorithms without contradicting their status as computationally Bayesian. In this way, the explicit-Bayes hypothesis is intended as a complement to rather than a replacement of the view that cognition is Bayesian at the computational level.

Testing the explicit-Bayes hypothesis

The explicit-Bayes hypothesis represents a specific, falsifiable hypothesis: within a given model space, the best-fitting model either explicitly uses Bayes' rule (supporting the explicit-Bayes hypothesis for that task) or does not (counting against the explicit-Bayes hypothesis for that task). To illustrate, consider two models of how people judge whether a briefly flashed stimulus belongs to category A or B. An explicit-Bayes model would make this decision by computing posterior probabilities from the likelihood of the internal evidence under each category together with

the prior probability of each category. By contrast, a non-explicit-Bayes model might make the decision by comparing the strength of internal evidence to a decision threshold without explicitly computing sensory likelihoods or using Bayes' rule. In this example, the explicit-Bayes hypothesis would be supported for this task if models explicitly applying Bayes' rule outperform models that simply compare evidence to a threshold; the opposite result would count against the explicit-Bayes hypothesis.

Models that do or do not explicitly use Bayes' rule might produce identical outputs: in such situations, the explicit-Bayes hypothesis would be neither supported nor rejected until further data are collected to distinguish between these models. Whereas a single experiment can only support or reject the explicit-Bayes hypothesis for a given task, repeated tests across related tasks can reveal whether the hypothesis generalizes across broader domains of cognition.

A practical difficulty for testing the explicit-Bayes hypothesis is that for any given task, one could formulate many explicit-Bayes and non-explicit-Bayes models. Nevertheless, this challenge is manageable with the right experimental design: choosing a behaviour that is simple enough that the set of possible models is tractable, while ensuring sufficient complexity in the data (such as a high number of conditions) to make different models identifiable. For example, memory for real-world events is too complex and high-dimensional and leads to too many possible models, whereas testing whether Gabor patches of a fixed contrast are tilted left or right might be too constrained and low-dimensional and lead to models that are not identifiable. However, a Gabor orientation task coupled with a factorial design featuring four different priors, four reward structures and four Gabor contrasts creates a useful setting. This problem is low-dimensional, only a few models are possible, and with 64 conditions, these models are identifiable. More generally, because explicit-Bayes and non-explicit-Bayes models can sometimes mimic each other over a restricted set of conditions, a broad range of manipulations makes it easier to distinguish between these two classes of model.

One particularly promising domain for testing the explicit-Bayes hypothesis is confidence in perceptual decisions. Perception represents an especially tractable domain relative to higher-level domains such as memory or reasoning, whereas confidence judgments are useful because participants are often explicitly or implicitly instructed to report probabilities — precisely the quantities that Bayes' rule computes. The explicit-Bayes hypothesis in the context of confidence is known as the 'Bayesian confidence hypothesis' and has been tested numerous times[6,7]. The field has also identified a major non-explicit-Bayes alternative: that people make confidence judgments based on the strength of sensory evidence, without applying Bayes' rule[8]. Comparisons of explicit-Bayes models and non-explicit-Bayes models have found support for the latter models[8–10]. Although preliminary, these findings suggest that the underlying algorithms for confidence in perceptual decisions do not explicitly use Bayes' rule. More broadly, this work demonstrates the feasibility of testing the explicit-Bayes hypothesis within a specific cognitive domain.

Outlook

The explicit-Bayes hypothesis is a specific hypothesis about the nature of the algorithms used to solve cognitive tasks. In contrast to the Bayesian approach that considers the computational level, this hypothesis is readily falsifiable and its formulation at the algorithmic level connects it more tightly to other modelling work in cognitive science. As a programme of research, testing this hypothesis transcends specific subdisciplines and brings insight into the mechanisms that underlie cognition.

The effort to determine whether cognitive algorithms explicitly use Bayes' rule complements rather than replaces the more traditional effort to use Bayesian inference at the computational level as a tool for asking why behaviour is the way it is. Indeed, if the explicit-Bayes hypothesis were to repeatedly fail across many domains, this pattern would not refute computational-level Bayesian models. Instead, it would constrain how they are interpreted: they would be more naturally seen as normative or abstract characterizations of cognition rather than as descriptions of the algorithms that cognition uses. In this way, testing the explicit-Bayes hypothesis can help clarify not only the nature of cognitive algorithms but also the explanatory role of Bayesian models at the computational level.

Kai Xue ✉ **& Dobromir Rahnev**
School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA.
✉e-mail: kxue33@gatech.edu

Published online: 29 April 2026

References

1. Jones, M. & Love, B. C. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* **34**, 169–188 (2011).
2. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 2010).
3. Griffiths, T. L., Chater, N., Norris, D. & Pouget, A. How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychol. Bull.* **138**, 415–422 (2012).
4. Gigerenzer, G. in *Blackwell Handbook of Judgment and Decision Making* (eds Koehler, D. J. & Harvey, N.) 62–88 (Wiley, 2004).
5. Parpart, P., Jones, M. & Love, B. C. Heuristics as Bayesian inference under extreme priors. *Cognit. Psychol.* **102**, 127–144 (2018).
6. Li, H.-H. & Ma, W. J. Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* **11**, 2004 (2020).
7. Aitchison, L., Bang, D., Bahrami, B. & Latham, P. E. Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* **11**, e1004519 (2015).
8. Locke, S. M., Landy, M. S. & Mamassian, P. Suprathreshold perceptual decisions constrain models of confidence. *PLoS Comput. Biol.* **18**, e1010318 (2022).
9. Xue, K., Shekhar, M. & Rahnev, D. Uncovering internal evidence representations reveals the nature of confidence computation in multi-alternative perceptual decision making. Preprint at https://doi.org/10.31234/osf.io/67ufv_v2 (2025).
10. Xue, K., Shekhar, M. & Rahnev, D. Challenging the Bayesian confidence hypothesis in perceptual decision-making. *Proc. Natl Acad. Sci.* **121**, e2410487121 (2024).

Competing interests

The authors declare no competing interests.