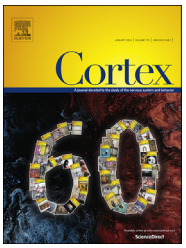
CORTEX 170 (2024) 69–79

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex



Viewpoint

The utilitarian brain: Moving beyond the Free Energy Principle

Babak Hemmatian [a], *Lav R. Varshney* [a,b], *Frederick Pi* [c] *and Aron K. Barbey* [a,d,*]

[a] *Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign, USA*
[b] *Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, USA*
[c] *Department of Cognitive Science, University of California San Diego, USA*
[d] *Center for Brain, Biology and Behavior, University of Nebraska Lincoln, USA*

ARTICLE INFO

Article history:
Received 17 October 2023
Reviewed 23 November 2023
Revised 28 November 2023
Accepted 28 November 2023
Action editor Gus Buchtel
Published online 7 December 2023

Keywords:
Free Energy Principle
Subjective utility
Extended cognition
Decision-making
Cognitive neuroscience
Bayesian Brain Hypothesis

ABSTRACT

The Free Energy Principle (FEP) is a normative computational framework for iterative reduction of prediction error and uncertainty through perception–intervention cycles that has been presented as a potential unifying theory of all brain functions (Friston, 2006). Any theory hoping to unify the brain sciences must be able to explain the mechanisms of decision-making, an important cognitive faculty, without the addition of independent, irreducible notions. This challenge has been accepted by several proponents of the FEP (Friston, 2010; Gershman, 2019). We evaluate attempts to reduce decision-making to the FEP, using Lucas' (2005) meta-theory of the brain's contextual constraints as a guidepost. We find reductive variants of the FEP for decision-making unable to explain behavior in certain types of diagnostic, predictive, and multi-armed bandit tasks. We trace the shortcomings to the core theory's lack of an adequate notion of subjective preference or "utility", a concept central to decision-making and grounded in the brain's biological reality. We argue that any attempts to fully reduce utility to the FEP would require unrealistic assumptions, making the principle an unlikely candidate for unifying brain science. We suggest that researchers instead attempt to identify contexts in which either informational or independent reward constraints predominate, delimiting the FEP's area of applicability. To encourage this type of research, we propose a two-factor formal framework that can subsume any FEP model and allows experimenters to compare the contributions of informational versus reward constraints to behavior.

* *Corresponding author*. Center for Brain, Biology and Behavior, University of Nebraska Lincoln, USA.
E-mail address: abarbey2@unl.edu (A.K. Barbey).

https://doi.org/10.1016/j.cortex.2023.11.013
0010-9452/

1. Introduction: For the love of a paradigm

Establishing a unifying theory, a "paradigm" in Kuhnian history of science, is a Holy Grail of any scientific discipline (Kuhn, 2012). As demonstrated by the success of atomic and evolutionary theories in physical and biological sciences, it is possible to have such unifying theories that transform piles of "incommensurable" evidence into additive parts of a bigger whole, not only guiding but also encouraging research in relevant domains. In the absence of a similar grand theory, the psychological and brain sciences remain in a "pre-paradigmatic" state of unenviable confusion and infighting, though not for a lack of trying. The radical behaviorism most eloquently championed by Skinner (2002) was one attempt to establish an all-encompassing paradigm for psychological science. It failed because its proposed machinery could not begin to explain the entirety of human behavior (Tolman, 1948), and the notion of reward it employed was deemed too shapeshifting to allow falsifiability (Popper, 1972, p. 295).

But, it is argued, the Free Energy Principle (FEP) could potentially fill the same ambitious role within the brain sciences (Friston, 2010; Friston et al., 2012).[1] While the computational machinery associated with the notion is complicated, the core ideas are as intuitive as those of any good unifying principle: Humans are driven to reduce the error and uncertainty ("Free Energy") in their predictions of environmental outcomes by building mental models that correspond to their experiences, tweaking the models as bidden by informational feedback. "Active inference" presents an important complement to this type of corrective "sensing" in reducing Free Energy, as it guides an agent's actions to remake the environment in the image of their internal representations (Clark, 2015). It has been suggested that the principle provides the essence behind the turning of our neuronal cogs,[2] the guiding principle of our phylogenetic neural evolution,[3] and even determines the strategies we employ to solve problems.[4] To perform the Principle's computational edicts normatively requires solving intractable equations. The "Bayesian brain", however, could use approximate Bayesian inference to find practical solutions up to a given error bound (see Friston et al., 2006, for a detailed discussion), as satisfying an answer as a limited biological entity could hope for. Thus, a brain focused on reducing Free Energy could claim (approximate) Bayesian optimality.

Years have passed since the most impactful statement of the principle as a paradigm candidate for the brain sciences was published (Friston, 2010). In the intervening years, the challenge to apply the FEP throughout the brain-related disciplines has been answered by ingenious attempts to explain every aspect of higher cognition using this framework, from conversing in natural language (Friston & Frith, 2015) to the very nature of general intelligence (Ashton Smith, 2023).

For the attempts at fashioning the Principle into a paradigm for the brain sciences to succeed, it must provide a unifying account of the constraints bearing on our brains. All our major cognitive functions, our non-idiosyncratic abilities, must be reducible to attempts at minimizing prediction error or uncertainty, be it through structured perception or iterative actions on the environment. As Lucas (2005) laid out in a *Cortex* theoretical article, the constraints to explain would encompass the intertwined biological, neural, and cultural contexts in which cognition is embedded. He proclaimed that no extant theory of the brain satisfactorily fills this difficult role, but his summary of the field came before the most impactful call to action regarding the FEP in brain science (Friston, 2010), as well as much of the work applying the framework to higher cognition (e.g., Friston & Frith, 2015; Gershman, 2019).

We therefore use the meta-theory proposed by Lucas (2005) as a guidepost to evaluate how well the FEP has fared as a potential paradigm for the brain sciences. To make the problem more tractable, we focus on one domain rather than the entirety of human thought and behavior: classic problems within decision-making research and some of the traditional experimental paradigms used to model them (Keren & Wu, 2015). The FEP's proponents clearly consider this domain within the theory's scope, as it figures prominently in both early articles proclaiming the Principle's unifying potential (Friston, 2010), and more recent extensions of the theory (e.g., Ashton Smith, 2023; Gershman, 2019; Sprevak & Smith, 2023; Schwartenbeck et al., 2013). But regardless of whether it has been a focus of FEP-inspired research, decision-making makes for such an integral aspect of higher cognition that any theory claiming to unify the brain sciences would need to explain it adequately without the addition of independent, irreducible notions.

[1] "This diversity allows the [Free Energy] principle to account for many aspects of brain structure and function and lends it the potential to unify different perspectives on how the brain works." (Friston, 2010, p. 127). A more definitive statement is found in Friston et al. (2012), despite conceding that certain "idiosyncratic" aspects of cognition may not be explainable by the FEP: "Contrariwise, the greatest virtue of the free-energy framework, it seems to me, is that it reveals the underlying unity beneath that superficially heterogeneous array of ploys and policies, displaying bodily form, biomechanics, learning, niche-construction, perception, and action as manifestations of a single ongoing adaptive imperative to reduce informational surprise. The resulting unified model of brains, bodies, and active, environmentally embedded agents seems to me to be one of the most exciting new developments in the ancient quest to understand mind and its place in nature."

[2] "Not only do hierarchical models have a key role in statistics (for example, random effects and parametric empirical Bayes models), they may also be used by the brain, given the hierarchical arrangement of cortical sensory areas" (Friston, 2010, p. 129) or "the free-energy principle entails the Bayesian brain hypothesis" (Friston, 2010, p. 130). See also the discussion of sufficient statistics needed for Bayesian inference on page 130 where it is simply assumed that "the brain encodes these statistics".

[3] "… agents move through a succession of states that have acquired value to access states (rewards) with genetically specified innate value. … So how does this relate to the optimization of free energy? The answer is simple: value is inversely proportional to surprise, in the sense that the probability of a phenotype being in a particular state increases with the value of that state. Furthermore, the evolutionary value of a phenotype is the negative surprise averaged over all the states …" (Friston, 2010, p. 133).

[4] "More generally, it shows how rewards and goals can be considered as prior expectations that an action is obliged to fulfil." (Friston, 2010, p. 134).

To foreshadow, we identify classes of decision-making problems with well-documented accompanying preferences and strategies that the Principle cannot explain without FEP-independent extensions. We trace these failures to the lack of an adequate notion of utility, i.e., subjective valuation, in the theory. We then examine several attempts at extending the FEP to incorporate this cornerstone of decision-making research in ways that fully reduce it to error/uncertainty reduction (e.g., Friston, 2010, pp. 133–135). We present an array of experimental and everyday decision-making contexts in which these extensions fail and discuss why no similarly reductive attempt is likely to succeed in the future.

The history of science is awash with irony. A paradigm candidate proposing the reduction of reward to cognitive variables may run into the same broad hurdles that its precursor reducing cognition to reward found fatal: A foundational notion that is difficult to falsify without reinforcement using independent concepts (Popper, 1972, p. 295), and inadequate machinery for dealing with certain dimensions of human thought and behavior (Tolman, 1948). What are the prospects of the FEP in brain science considering these hurdles? After the failure of early cognitive research to present a unifying paradigm, many researchers settled for a more circumscribed goal, that of finding an ever-present principle that crosses phylo-, onto-, and epigenetic boundaries on top of disciplinary ones, to serve as a *component* of a future paradigm. Shepard's "Universal Law of Generalization" (Shepard, 1987) provides a prominent example of this approach. Can the FEP satisfy the same role, as an ever-present, albeit incomplete, component of the brain's processing?

We discuss decision contexts where the contributions of the FEP appear minimal, if at all present. To preview our closing argument, this is because prediction error and uncertainty matter to decisions *insofar as* they relate to an agent's goals, hence their valuation of different sensory states and actions. While utility appears irreducible to error and uncertainty reduction, any FEP-derived model could in principle be incorporated into a utility-dependent model of choice valuation.[5] The connections between reinforcement and deep learning research on one hand and brain research on the other provide a unique opportunity for supercharging cross-disciplinary research, if neuroscientific models do not mischaracterize utility using needlessly reductive approa ches.

We recognize that not all extensions of the FEP to decision-making have been reductive. For example, Sprevak and Smith (2023) developed a non-reductive model of decision-making that incorporates a prior preference distribution, rather than relying solely on the FEP. Furthermore, Smith et al. (2022) established a model of well-being that incorporates the Principle but uses subjective preference as a foundational input. Despite differences with what we propose in this article, we believe such non-reductive models provide a promising path forward in decision-making research. However, they do not have the same implications for unifying the brain sciences that the reductive approaches do: By design, they recognize that the Principle's notions of error and uncertainty do not suffice to account for behavior in some contexts and may be less relevant in others. As such, they fall outside the scope of our current critique.

If the FEP is neither all-encompassing nor ever-present in human decision-making, it would not be a unifying principle for the brain sciences, let alone a paradigm. It would simply be yet another pre-paradigmatic theory among the many produced by brain researchers, each with its own delimited, even if broad, domain of applicability. While this would be a humbling, unfortunate outcome, it is one that all but a handful of theories in science have faced, an eventuality that does not automatically deny a framework its value as an instrument of disciplinary research.

Rather than unsuccessful attempts to reduce every brain function to prediction error and uncertainty reduction, it would be more productive to identify contexts in which the FEP's considerations most strongly determine behavior. Beyond supporting the Principle's application to domains where it is most useful, such an approach would allow its composition, if and where needed, with other fundamental notions, in hopes of providing the brain sciences with a unifying theory in the future.

To advance the discourse along these lines and considering our preceding discussion of independent subjective preference, we subsume the FEP into a preliminary two-factor computational model that separates its informational concerns from irreducible aspects of utility estimation. The relative contribution of the two factors is controlled by separate parameters that may be empirically estimated. Importantly, such a model can still accommodate normative, Bayesian modeling, while also allowing the development of descriptive computational theories. In the interest of brevity, we leave to a future article the full details of how the parameters would be estimated in such a model and how to experimentally test its predictions. We give the interested reader a broad equation and pointers to relevant experimental literatures for now. We end the article by briefly discussing what ramifications this more contextualized view of the brain, in line with Lucas' suggestions in his 2005 *Cortex* article, has for the next 60 years of studying the human cortex.

2. Free Energy versus utility in decision-making

Decision-making is a central human faculty. For the FEP to serve as a unifying principle for the brain, it needs to address its processes and outcomes without a need for additional, irreducible notions. Valuation is at the core of decision-making, reflected in the presumption of subjective utility's existence across the behavioral decision sciences (Barberà et al., 2004). For the FEP to unify the brain sciences, decision goals and their mapping to options according to this valuation process must be reducible to error and/or uncertainty reduction in ways reliably predicted by the Principle. That the

[5] In fact, the Principle's informational considerations are arguably already reflected in computational cognitive neuroscience of decision-making, where approximate Bayesian inference is performed nowadays using neural networks (Fengler et al., 2021). It is noteworthy that reinforcement learning in such models prefers the reduction of uncertainty and error without explicitly using the FEP, as less uncertain outcomes are often important for the decisions of a rewards-maximizing agent.

theory's proponents consider short- and long-term decision-making within the scope of the FEP's unifying power is reflected in several attempts to explain away choice valuation as a function of the Principle's core equations, including within the most impactful call to action for exploring the theory's potential in the brain sciences (Friston, 2010; Gershman, 2019).

The idea that the FEP obviates the need for independent notions of value, reward, or utility features prominently in Friston et al. (2009, 2014), but is most broadly stated in an article titled "The free-energy principle: a unified brain theory?" (Friston, 2010). There, it is argued that value could simply be the complement of surprise, while a goal could be considered a state an agent expects to spend much of its time in. How did this valuation of the most expected outcomes develop? The explanation is that natural selection pressured neural ensembles capable of associative learning (Hebb, 2005; Von Der Malsburg, 1994), leading to the development of attractor states that an agent predicts and is simultaneously drawn towards. More recently, the same notion was used to explain goal-directed behavior as key to general intelligence, building an account of decision-making most broadly construed based on choosing the option with the least expected prediction error or uncertainty (Ashton Smith, 2023).

If we take "the most expected option" to mean the option with the highest *a priori* probability, as is the common sense of the term and seemingly the definition used in Friston (2010, p. 133), it is difficult to see how our account would provide a compelling case for two classes of reasoning problems and their associated decision rules. We look at each of these classes briefly in turn.

In diagnostic reasoning an agent reasons backwards from an effect to its cause, trying to identify what could have created an observed attribute. The most intuitive everyday example of this type of task is clinical testing. Consider a lab test devised to identify cancer. As in all decision-making contexts, there are two possible ways in which the test could lead us astray: There could be no cancer in a sample while the test flags it as cancerous (Type 1 error), or the sample could be cancerous while the test identifies no abnormality (Type 2 error). It is not difficult to see why cancer tests should be optimized to minimize Type 2 error even at the expense of increasing false positive rates: A false positive outcome exacts a miniscule cost, as further testing would dispel any worries; while missing a present cancer could prove fatal to a patient. In this context, a cancerous sample is not the most expected outcome. In fact, if the test is routine, it has very low probability. However, the negative utility of a false negative is so immense that it overpowers the probabilistic constraints and guides our decision about which test to employ. The same is not true of every diagnostic reasoning context. In research, for instance, it is customary to cap Type 1 error rates at .05 while Type 2 error rates are allowed to get as high as .2. The reasoning is that a fake positive result can lead to much greater waste of research potential than a missed positive finding. This adaptive adjustment of error rates in line with the utility of less likely outcomes is an example of the immediate physical context constraining brain processing in a structured manner (Lucas, 2005).

Predictive reasoning is the opposite of diagnostic thinking, reasoning forwards from a cause to determine what effects it might have. Nonetheless, it poses similar dilemmas for identifying utility with expectation. Consider instrumental predictive reasoning, or identifying actions one should take to obtain a desired outcome. The pursuit of political ideals often involves this type of processing.[6] Many throughout history have been willing to die for their visions of a better society even when the fight has been hopeless, i.e., when the probability of achieving the goal state approached (or even reached) zero. In this case, the positive utility of even imagining the political ideal being implemented is so great that it can overpower even the deep-seated self-preservation instinct. The ideal has that effect *without* requiring that the goal state be likely, just the knowledge that it is *desired*. Like diagnostic reasoning, predictive problems need not be about extreme utility. A more mundane example would be determining how much effort to spend on preparing for a test. The adjustment of actions based on the utility of what they support provides an interface for the intermediate cultural context to constrain the brain's decision processes (Lucas, 2005).

One possible way to avoid these issues while maintaining that utility boils down to expectation in the traditional FEP sense is by changing the intuitive meaning of the latter term: Perhaps "not having active cancer", "being a good fighter in the path of one's ideals" or "having gotten good grades" are the awkwardly defined yet imaginable states that our evolution has led us to "expect" being in most of the time. There are two major problems with this idea. The first is that one's desire towards such states might still differ strongly from one's expectation. We might consider it most likely that we will fail an exam yet be motivated to study on the off chance that the desired, less likely outcome could happen. The second problem with this framing is that it would make the notion of expectation so broad and so vague as to render the FEP's application to such decision-making contexts all but untestable. That would obviate any empirical benefits from the Principle's adoption as a paradigm for the domain.

In contrast, the simplest FEP-independent notions of utility, i.e., the value of an outcome or an event expressed in terms of an individual's personal judgment or degree of satisfaction (Oxford University Press, 2009), can easily account for behavior in the cases mentioned. Where does this valuation come from? One could simply trace it to pleasure and pain as shaped by evolutionary pressures (Hagen et al., 2012). Note that adopting a non-reductive utility-based approach to reasoning and consequently decision-making does not necessitate abandoning the Bayesian Brain Hypothesis. Bayesian accounts of diagnostic and predictive reasoning have long been developed (Fernbach et al., 2011), and expected utility approaches allow for the incorporation of risk in translating them into decision rules (Fishburn, 1981).

If goals are not most fruitfully modeled as the most expected outcomes, is there an alternative way of reducing utility to Free Energy minimization? The most plausible formal candidate for this role is discussed by Gershman (2019). His account considers utility maximization as the goal of

[6] The FEP has been used to explain communication and hermeneutics (Friston & Frith, 2015). Therefore, the theory's proponents seem to consider abstract concepts like a political ideal within the scope of its unifying potential.

decision-making but casts reward as information gain. This essentially has the effect of turning the agent's focus away from prediction error reduction towards uncertainty reduction. In other words, it suggests that agents are compelled to take actions that teach them the most about the potential states of their environment.

That the well-studied phenomenon of "motivated forgetting" exists and has been studied in the brain (Anderson & Hanslmayr, 2014) should immediately give us pause in considering this a complete account of human behavior. In cases of motivated forgetting, information is actively removed from memory, such that uncertainty (and plausibly often error) is increased. This is done in ways that have positive utility for the agent, for instance by allowing them not to ruminate on expected experiences that are unpleasant. But for the sake of the argument, let us consider such phenomena as edge cases and look at the error and uncertainty reduction aspects of the FEP in more canonical decision-making contexts.

In many such situations, the two elements of the FEP conflict, with no clear way of choosing between the two without an independent notion. Consider, for example, a classic multi-armed bandit problem (Gittins, 1989). In this class of problems, a fixed, limited set of resources must be allocated between competing, alternative choices, when each choice's properties are only partially known at the time of allocation and may become better understood as time passes or by allocating resources to the choice. A common example is choosing between different restaurants. Consider two options: a familiar restaurant with decent food and a newly opened restaurant that is an unknown quantity. For simplicity, assume that the new place has a 50% chance of being terrible and a 50% chance of being great in the agent's mind. In this context, an information gain goal would support going to the new restaurant. However, unexpectedly eating terrible food is a prediction error the agent would also want to avoid. Wanting to minimize that error would lead to choosing the safe option. It is not clear which goal should be more prominent according to the FEP. One possibility would be to define the expected state the agent is drawn towards less intuitively as "having eaten good food". This would encourage exploration or exploitation based on which option is more likely to bring that state about. But the key term in such a characterization would be "good", i.e., the subjective valuation or utility that appears indispensable.

Knowing the subjective utility helps us explain behavior in this and many other multi-armed bandit problems. Depending on the agent's risk-aversion (Edwards, 1996), if the known restaurant is only decent, i.e., the difference between its expected utility and the perceived maximum available in one's environment is large, agents may "explore" by visiting the new restaurant. But if they feel that the known location "satisfices" as an option, i.e., surpasses their personal threshold for how good a meal should be, there is less impetus for exploration. Such utility-based solutions to bandit problems are so readily formalizable that reinforcement learning-based approaches to finding them are used in many applied settings (Bouneffouf & Rish, 2019).

In less mundane contexts, an irreducible notion of utility provides even more clear-cut predictions for whether uncertainty reduction would be pursued. Consider someone who sees the vague silhouette of a large animal at night in the wilderness. It could be a harmless grazing deer or a vicious leopard. In such a situation, the negative utility of being within the reach of a potentially dangerous animal outweighs any decision-relevant insights to be gained from reducing uncertainty about the shadow's identity. This means that most individuals would simply keep their distance and let go of their information gain goals. A contrasting situation where information gain aligns more with utility is in the Twenty Questions game, information-theoretic accounts of human behavior in which have been developed (Nelson, 2005).

Even when focusing on highly similar decision contexts, the preference for prediction error versus uncertainty reduction can dynamically change in ways predicted by subjective preference. For students practicing in preparation for an exam, information gain may be more important: They would want to focus on problems the answers to which appear most uncertain (although even then the incentive structure matters for determining one's strategy; Oxoby, 2009). But the relative importance of information gain flips for the test itself: Reduction of prediction error would be more important than "trying out" uncertain strategies to learn more about the topic while one's grade hangs in the balance. The distinction can be explained by noticing that much less negative utility is attached to making mistakes during a practice than in an actual test. Note that "having obtained good grades" or "remaining healthy" may be highly unlikely outcomes, but that does not detract from their desirability.

It is difficult to see how *any* notion of utility that relies solely on prediction error or uncertainty reduction would fully explain humans' highly adaptive behavior in response to these diagnostic, predictive and bandit problems. Responses to these tasks are linked to an unavoidable property of natural cognitive agents: One may desire states that not only differ from those of the environment or the most likely outcomes, but that could even be unachievable in principle. This is simply a consequence of our brains maintaining two interconnected yet different systems, one for representing real and imaginable worlds, and another for evaluating the utility of various states. This distinction is reflected in the neuroscience of decision-making, where representation and evaluation's neural underpinnings are distinctly characterized (Gold & Shadlen, 2007; Rangel et al., 2008).

A side effect of this separation is that in some decision-making contexts neither prediction error nor uncertainty reduction may be important: Consider a situation where an agent is forced to perform a decision-making task, or the reward does not depend on the quality of responses. In this situation, there is little motivation to find the right answers, as the utility is uniformly distributed across each trial's options. A subject might therefore choose the least effortful action of randomly responding instead of trying to reduce uncertainty or prediction error. Even the indifference between options in this example is explainable using utility. The effectiveness of manipulating the utility distribution is reflected in the fact that researchers often add a bonus to a subject's baseline compensation based on performance to encourage accurate response. In comparative neuroscience studies with monkeys, this takes the form of sweet nectars offered through a tube

(e.g., Montague & Berns, 2002). In many human experiments, a random trial is chosen to determine a monetary bonus (e.g., Tomov et al., 2020). These interventions tip the expected utility balance in favor of minimizing the subjects' Free Energy in the task, giving them a reason to focus their efforts on prediction error or uncertainty.

This example not only denies the FEP its ubiquity across brain functions in general and decision-making contexts in particular, but it also highlights a broader point about when its considerations would be relevant to a task: Prediction error and uncertainty reduction matter to agents *insofar as* they relate to subjective utility. That is why focusing on long-term optimization rather than short-term decisions (as we have in this section, in contrast with Friston, 2010) would not save the reductive FEP-based accounts of decision-making: Agents are driven to optimize their long-term utility as well as their short-term rewards. If any part of the utility to be accrued in the long-term does not depend on the FEP's considerations, that undermines the Principle's ability to explain behavior with broader horizons.

To summarize, while it is unlikely that utility can be reduced to Free Energy minimization, the only characterization of Free Energy that could hope to account for human decision-making would be one complemented by irreducible notions of subjective preference. If the FEP is less relevant to some decision-making contexts and cannot independently explain human behavior in classes of heavily studied problems, it falls short of being a unifying principle, let alone a paradigmatic theory to subsume all of brain science. The issues we raised in this section with accounts of choice valuation that rely solely on the FEP appear profound. Therefore, we do not find further pursuit of reductive research projects at the intersection of the FEP and decision-making (e.g., Friston, 2010; Gershman, 2019) helpful.

3. Free Energy and utility in decision-making

If we concede that the FEP is an unlikely candidate for a unifying theory of the brain, what would the prospect be for moving towards a paradigmatic brain science? One possibility is that there will never be a unifying theory. Although the optimization approach to biology has been fruitful in explaining why brains developed certain properties (Varshney et al., 2016), some describe the mind-brain system as a kludge—a quick-and-dirty solution that is clumsy, inelegant, inefficient, difficult to extend, and hard to maintain—due to the limitations of evolution (Marcus, 2009) and perhaps of development (Witvliet, et al., 2021).

But the brain scientist's task would remain the same regardless of whether a unifying theory can be developed: Delineating each framework's area of applicability. This is because even the broadest paradigms of the physical sciences still have delimited domains of explanatory power. If our brains are alternatively like Swiss army knives, a set of ill-aligned tools that are each pulled out according to situational demands, we should still identify the contexts in which a framework's edicts are most prominent. If we find a way of formally combining the various tools, we could still develop a modular model of the brain in its totality which, though not nearly as elegant as a single theory, would have far greater explanatory power than would otherwise be available to us.

If we accept the delineation of theories' domains as our task, what could be the area where the FEP is most helpful without independent notions? When the learning of generative models aligns with an agent's utility considerations, the FEP explains behavior well. One example is trial-and-error in causal learning, where active inference is arguably used to iteratively reduce uncertainty about effects by using prediction error as a teaching signal (Friston et al., 2009).

To make use of the FEP's explanatory value in such contexts while maintaining greater range for explaining brain function, we could reframe subjective utility such that it isolates the influence of FEP's information optimization goals, compares it with irreducible aspects of subjective preference and allows for the selective activation of either element based on situational demands.[7] Such an extension would make it easier to identify the Principle's domain of applicability, therefore integrating FEP findings with exciting developments in the formal modeling of higher-level cognition and bringing us closer than ever to the goal of providing a unifying computational account of the brain. There are already cognitive theories that combine notions of uncertainty reduction with the optimization of utility in domains as complex as social learning (FeldmanHall & Nassar, 2021) and moral reasoning (Cushman et al., 2017). Such models would benefit from further integration of their normative aspects with the FEP if the necessary and superfluous contributions of the Principle to utility estimation are distinguished. Even if the Bayesian Brain Hypothesis proves to be false in certain domains and individuals are found to deviate systematically from the normative edicts, a clearer quantification of those deviations using normative models like the FEP would set the stage for better descriptive accounts.

In what follows, we provide a preliminary version of what a utility model could look like that directly compares the impact of FEP as a module with that of irreducible subjective preference. Because prediction error is the main determinant of actions in the canonical version of the FEP (Friston, 2010), we focus on this element in the following discussion. But similar accounts could be easily developed for alternative, more sophisticated notions of prediction error as well as versions of the FEP that focus on uncertainty reduction (e.g., Gershman, 2019) or combine error and uncertainty in a structured manner.

Let s stand for the decision context that has produced certain sensations in the agent. In the context of choosing a restaurant, this could be the agent's level of hunger. Let the vector $A = \{a_1, \ldots, a_n\}$ denote the possible actions the agent could take in light of the sensations. The action could be the agent going to Restaurant One or Restaurant Two. We could represent the possible outcomes as the vector $Y = \{y_1, \ldots, y_n\}$, in this case how filling the food is, i.e., how good of a solution it

[7] Alternatively, one could combine utility with the FEP by using subjective preferences as constraints on a model optimizing Free Energy. An example is the model of holistic wellbeing developed by Smith et al. (2022). Whether using utility as a constraint in an FEP model produces better or worse explanations for behavior than vice versa, awaits further research.

provides to the current problem sensation. The agent's valuation of a given action in the context of the sensations could then be determined using the sum of two terms, one representing prediction error for the action outcomes couched in information-theoretic terms, and the other the utility of outcomes relative to the status quo. In our example, Restaurant One may have been visited many times, in which case it would be associated with very low expected prediction error: Our agent would know rather accurately how filling their food will be and is therefore unlikely to heavily update the learned distributions based on outcomes. The utility, however, has to do with more than this knowledge. Perhaps our agents would expect the portion sizes at Restaurant One to be too large for their current level of hunger. This would encourage them to try Restaurant Two instead, because an overfull belly has a negative utility. Of course, whether they would prefer Restaurant Two depends on their expectation of how large the portion sizes will be there. If they have never been to Restaurant Two, they may think of the average portion size at a restaurant of that type. An intuitive characterization of either action's valuation is given in Equation (1)[8]:

$$V(a_i|s) = \alpha(s)\left(\sum_y U(y|s,a_i)p(y|s,a_i) - U(s)\right) - \beta(s)\left(D[p(y|s,a_i) \| p(y|s)]\right) \quad (1)$$

An agent considers through the first term each possible outcome of an action in terms of its expected utility given the sensations. Importantly, the expected utility after the action is considered relative to the status quo, $U(s)$: There is no point in taking effortful steps if the utility remains the same.

For the second term, the expected difference between the outcome probability distribution before and after the action is calculated using Kullback–Leibler Divergence, an entropy-based measure of prediction error.[9] To encourage the minimization of this difference, negative value is assigned to the divergence.

Here, $\alpha(s)$ and $\beta(s)$ would be the context-specific and empirically derived normalized weights that determine the relative prominence of prediction error reduction and independent utility maximization in the overall valuation scheme. This form allows the FEP to be the sole consideration in certain cases, or for it to be completely abandoned in favor of utility maximization in others. We anticipate the weight of irreducible subjective preference to be greater in our running example of choosing a restaurant. In other situations, however, FEP-based considerations may hold greater independent weight. For instance, if we are playing a Twenty Questions game, assuming that we are motivated to play well and win, we would probably choose questions that minimize our prediction error or maximize our information gain for the target category (Nelson, 2005). Since $\alpha(s)$ and $\beta(s)$ can vary based on context, they should be estimated empirically in each study. Once enough experimental evidence has accrued, the range of estimations across contexts can help identify domains where one term is most helpful for explanations, as well as places where a combination of information and reward optimization are necessary to explain behavior.

One context where such an interaction could feature prominently in overall action utility is during teaching. Ho et al. (2017) provide a compelling case in their experiments on navigating a grid world where some steps along the path to a rewarding destination are penalized. During learning trials, the players are encouraged to reduce their uncertainty and prediction error about where these "traps" are located, i.e., gather information about the distribution of (negative) rewards. The subjects are then told that other participants, also unaware of the traps' locations, would be shown records of their gameplay to learn how to succeed at the game. The subjects' behavior changes in these "teaching" trials. The best formal account for the subjects' responses is then provided by assuming that they are minimizing the uncertainty in the observer's expected distribution of rewards, a combination of Free Energy reduction and irreducible utility maximization. Ho et al.'s (2017) normative account is more focused on encoding-decoding accuracy than disentangling reward and information considerations. Nonetheless, it could serve as a helpful starting point for an analysis of the type proposed in this section.

Once the values of actions are determined using Equation (1), they can be transformed into normative choice probabilities using a variety of tools developed in the decision sciences (Barberà et al., 2004). This would allow researchers to compare human behavior directly with the predictions of a given utility model. A more normative transformation would be the widely used SoftMax function (Sutton & Barto, 2018). An alternative approach would be estimating context-dependent thresholds: Once an option's value surpasses that threshold, evaluation is ended, and the corresponding action is chosen. Which decision rule to use would depend on the researchers' hypothesis and the modeling's normative or descriptive goal. The key takeaway, however, is that translating the value estimations into behavioral predictions that can be fed into traditional statistical analyses for experiments would not be difficult.

Equation (1) is meant as an explanation at Marr's (2012) computational level of analysis: Representing the abstract problem that the agent wants to solve and the broad formal strategy for doing so. To transform it into an algorithmic account that, for instance, incorporates the limited resources of humans as boundedly rational agents (Lieder & Griffiths, 2020), a sequential ordering of behaviors needs to be mapped onto the computational framework. While the development and validation of such an account is beyond the scope of this viewpoint article, we present in Fig. 1 a simple decision

[8] Our proposed equation is very similar to the Lagrangian form of the novelty-utility trade-off in Varshney's (2019a) mathematical theory of creativity. The similarity is far from coincidental: Creativity is one domain where utility constraints ("can the creative solution aid the agent?") are intertwined with surprisal considerations in evaluating behavior (Guilford, 1967).

[9] The second term in Equation (1) is a contextualized variant of the objective minimized in variational Rate Distortion Theory (RDT). Incidentally, its optimization given an expected upper bound on distortion is given by a two-term characterization in the method of Lagrange Multipliers that is visually similar to Equation (1). The classic characterizations of variational RDT, however, contain no independent notion of utility. See Jakob and Gershman (2023) for an application of variational RDT to neural coding.

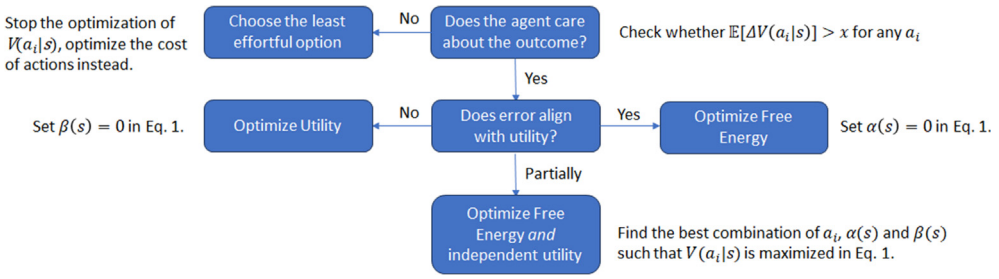


Fig. 1 – An example of how a model of choice valuation could be translated into an algorithmic computational theory that contrasts risk- and independent preference-based consideration. *x* in the top row is a contextually determined threshold for how much difference in expected option value is worth optimizing for. Utility and error are "partially" aligned when objective performance on the task matters for utility, but its value also reflects independent considerations (see text for examples). The first half of Eq. (1) can be used to model information-independent aspects of utility. Then $\alpha(\mathbf{s})$ and $\beta(\mathbf{s})$ may be optimized using grid search or through a context-sensitive heuristic.

tree to make the explanation level distinction clearer for readers. Each plain language element is accompanied by an example of how it could be formalized considering Equation (1).

Of course, to say that utility is indispensable to theories of cognition and can be empirically studied in brain and behavior does not mean its characterization would be easy in every context. Much like error and uncertainty, utility is a complex notion, its form shifting based on the decision parameters. More so than the FEP's purely informational considerations, it also has a subjective aspect that makes it more difficult to directly measure. In fact, it was this same subjectivity that prompted radical behaviorists like Skinner (2002) to resist conversion to cognitivism.

Still, there are clear-cut ways of grounding utility estimation in more objective measures. The largely dopaminergic network underlying reward-related perception and behavior is well-studied (Gold & Shadlen, 2007). Simple decision-making tasks and reinforcement learning-derived formalizations have proven successful at identifying the operating mechanisms of this network (e.g., Collins & Frank, 2014). By providing clear predictions about the brain regions and even the chemical processes underpinning choice valuation, the computational cognitive neuroscience of decision-making will be crucial for developing falsifiable, formalized accounts of utility estimation. The field's careful consideration of structural and functional brain constraints in these modeling frameworks also maps well onto the biological constraints in Lucas' (2005) meta-theory.

At a higher, more computational level of analysis (Marr, 2010), reinforcement learning algorithms have been successful at modeling behavior across a range of decision-making tasks, particularly when implemented within "drift diffusion models" (Fengler et al., 2021). These models have already been applied to explain the abstract, functional elements of cognition (Ratcliff et al., 2016). The functional focus of this field's research, removed from the implementational details of a brain or a computer's processing, maps well onto the abstract constraints in Lucas' (2005) meta-theory.

For researchers more interested in the FEP's insights about prediction, these research programs provide ample opportunities to perform integrative research, as they were built from the ground up with an awareness of uncertainty's role in decision-making. Such integrations could be more fruitfully developed if the reductive approach to utility (e.g., Friston, 2010; Friston et al., 2009, 2014) is not pursued any further.

4. Optimality versus efficiency in brain science

A major appeal of the FEP's computational machinery lies in the use of approximate Bayesian inference to estimate parameters in a manner that is optimal up to a given threshold. This allows normative models to be developed without adjusting the theory's fundamental equations, the claimed aligning of which with empirical data provides the basis for supporting the ambitious Bayesian Brain Hypothesis (Friston, 2010). Would the adoption of an independent preference concept mean the Hypothesis must be abandoned?

The answer is unequivocally no. Many ways of performing approximate inference are being developed that do away with the difficult-to-satisfy assumptions of classical Bayesian modeling while maintaining notions of optimality. Researchers in the computational decision sciences are now widely using neural networks for computations over models of utility and choice behavior not unlike those discussed in this article (Fengler et al., 2022). While many studies rely on drift-diffusion models of preference that we have not surveyed in this article (Ratcliff & McKoon, 2008), their use of universal function approximators means that the same deep learning frameworks could be applied to the estimation of any observable or latent variables in Equation (1), including in an approximately Bayes-optimal manner. While the training of neural networks is effort-intensive, it must be done for a given model-task pair only once. As such, the approach provides a promising avenue for future research on the neural underpinnings of behavior, which may bring us closer to a unifying theory of the mind-brain.

But researchers need not limit themselves to the predictions of the Bayesian Brain Hypothesis either. The functions that make up Equation (1) could be set up in a way that is most predictive of human behavior, which may turn out to deviate from optimality in reliable ways. How well such

models fit behavior could inform mechanistic theories of cognition. For instance, $U(y|s,a_i)$ in Equation (1) could be easily replaced with its Prospect Theory equivalent (Kahneman & Tversky, 1979) that incorporates risk aversion, while $p(y|s,a_i)$ could be adjusted to conform to the theory's predicted warped estimates of extreme probabilities.

We consider the descriptive focus particularly helpful for delineating the physical and cultural constraints in Lucas' (2005) meta-theory. Many judgments in complex social environments rely on heuristics activated by contextual cues (Hemmatian & Sloman, 2020), while time constraints make the use of normative strategies less likely even in canonical tasks (Payne et al., 1988). The fast-and-dirty rules-of-thumb offer a variety of cost-effective problem-solving approaches that are especially attractive to humans given their limited cognitive-behavioral capacity (Gigerenzer & Todd, 1999). Many heuristics have already been formalized and as such, can be incorporated into more encompassing computational models. Examples of such formalizations include domains like moral reasoning (Levine et al., 2020), suggesting that the same approach would be just as fruitful for explaining less-complex behaviors.

5. Summary and open questions

Unifying theories, or at least unifying principles that can serve as their components, are the Holy Grails of every science (Kuhn, 2012). We examined the performance of one proposed candidate for fulfilling this role in the brain sciences: the FEP (Friston et al., 2006; Friston, 2010). According to the theory, the cornerstone of brain processing is an attempt to reduce prediction error and uncertainty. Agents are driven to accomplish this through developing accurate models of the environment, but also by adapting it to better fit their internal representations (Clark, 2015). The challenge to reduce higher-level cognition to this Principle has been answered in a series of publications (e.g., Ashton Smith, 2023; Friston, 2010; Friston & Frith, 2015; Gershman, 2019). To evaluate the theory's potential for universal applicability across the brain sciences, we looked more closely at the reductive FEP-based models in a domain that has received much attention in recent years: human decision-making. We used Lucas' (2005) meta-theory of the embedded brain to frame our discussion of constraints on this cognitive function.

The bedrock of decision-making research is the subjective valuation of options, i.e., utility (Barberà et al., 2004). We argued that attempts to reduce this notion to FEP concepts, regardless of their focus on prediction error (e.g., Friston, 2010) or uncertainty reduction (e.g., Gershman, 2019), invariably fail to fully explain behavior in certain diagnostic, predictive and multi-armed bandit tasks. This is because the desirability of options does not solely depend on their expectedness, but also on irreducible subjective preference. That there is more to us than simply optimizing for surprise has implications far beyond the laboratory. For instance, when examining political decision-making in the context of today's social media, the utility of information must be considered on top of its surprise value to predict a message's persuasiveness (Varshney, 2019b).

To provide a stronger foundation for a future unifying perspective on the brain sciences, we presented a two-factor model of choice valuation that directly contrasts the informational considerations of the FEP with the independent utilitarian aspects of decision-making. We briefly discussed how the model's parameters may be estimated empirically using neural networks, as well as how more fine-grained algorithmic or implementational accounts may be developed. We believe the clear connections with network neuroscience (Barbey, 2018), reinforcement learning, and deep learning will prove crucial for the development of more comprehensive theories in this domain for years to come if needlessly reductive approaches to modeling are dropped.

We then reflected on our framework's implications for the Bayesian Brain Hypothesis (Hipólito & Kirchhoff, 2023). Our new framing of choice valuation allows for both normative and descriptive characterizations. The empirical estimation of variables using neural networks would also remain largely unchanged regardless of normative focus. As such, the utilitarian model can prove helpful in adjudicating between the accounts of decision-making behavior that rely on Bayesian principles and those that argue for systematic deviations from them. We ended by pointing to literatures that suggest each approach may be more useful for explaining human decision-making in certain tasks.

If we accept that utility must be characterized for us to have a full picture of human cognition, what would that mean for the future of brain science? There are three unavoidable properties of utility that make its characterization difficult. Firstly, it is only indirectly observed through behavior and must be estimated using latent variable analysis (cf., Barbey et al., 2021). Secondly, utility is constructed in each context based on the agent's goals, needs and task constraints. Thirdly, a complete image of human utility can only be derived by accepting its multi-dimensionality. Sen (1999) made a compelling case for the latter two properties in economics, but his critique applies just as well to cognition. While behavioral scientists often reduce utility to a scalar dollar value for ease of calculation, when thinking of complex notions like welfare, there is immense information loss associated with the dimensionality reduction. As such, for utility theories to remain true to their subjects beyond perceptual and economic game paradigms, the expansion of models to include different facets of preference in a contextual manner would be needed.

Fortunately, many approaches have been proposed for overcoming these complications in utility research, references to which are included throughout this article. Indeed, formal models of decision-making that combine the FEP with independent notions of subjective preference are already being developed (e.g., Smith et al., 2022). Beyond the existing models, promising connections exist in neighboring disciplines as we explored throughout this article. Therefore, even though we cannot yet rely on a single principle to explain all

behaviors and have much work to do perfecting our models of the irreducible subjective preference, there is no reason for despair. A unifying framework may yet be found.

CRediT author statement

Babak Hemmatian: Conceptualization, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration. **Lav R. Varshney**: Formal analysis, Writing – Review & Editing. **Frederick Pi**: Conceptualization, Formal analysis. **Aron K. Barbey**: Conceptualization, Writing – Review & Editing.

Funding

This work was supported by the Department of Defense, De fense Advanced Research Projects Activity (DARPA), via Con tract 2019-HR00111990067 to the University of Illinois Urbana-Champaign (PI: Aron K. Barbey). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Babak Hemmatian was supported with funding from the Beckman Institute for Advanced Science and Technology. Lav R. Varshney was supported by the National Science Foundation grant IIS-2123781.

REFERENCES

Anderson, M. C., & Hanslmayr, S. (2014). Neural mechanisms of motivated forgetting. *Trends in Cognitive Sciences, 18*(6), 279–292.

Ashton Smith, M. (2023, September 10). *The FEP+ Model of general intelligence*. https://doi.org/10.31234/osf.io/65gzh

Barberà, S., Hammond, P., & Seidl, C. (Eds.). (2004). *Handbook of utility theory (2 volumes)*. Springer Science & Business Media.

Barbey, A. K. (2018). Network neuroscience theory of human intelligence. *Trends in Cognitive Sciences, 22*, 8–20.

Barbey, A. K., Karama, S., & Haier, R. J. (2021). *The Cambridge handbook of intelligence and cognitive neuroscience*. Cambridge University Press.

Bouneffouf, D., & Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Collins, A. G., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review, 121*(3), 337.

Cushman, F., Kumar, V., & Railton, P. (2017). Moral learning: Psychological and philosophical perspectives. *Cognition, 167*, 1–10.

Edwards, K. D. (1996). Prospect theory: A literature review. *International Review of Financial Analysis, 5*(1), 19–38.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences, 25*(12), 1045–1057.

Fengler, A., Bera, K., Pedersen, M. L., & Frank, M. J. (2022). Beyond drift diffusion models: Fitting a broad class of decision and reinforcement learning models with HDDM. *Journal of Cognitive Neuroscience, 34*(10), 1780–1805.

Fengler, A., Govindarajan, L. N., Chen, T., & Frank, M. J. (2021). Likelihood approximation networks (LANs) for fast inference of simulation models in cognitive neuroscience. *Elife, 10*, Article e65074.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140*(2), 168.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision, 13*(2), 139–199.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Friston, K. J., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS One, 4*(7), Article e6421.

Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris, 100*(1–3), 70–87.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1655), Article 20130481.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 130*.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Gittins, J. C. (1989). *Multi-armed bandit allocation indices*. Chichester: John Wiley & Sons, Ltd.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535–574.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Hagen, E. H., Chater, N., Gallistel, C. R., Houston, A., Kacelnik, A., Kalenscher, T., Nettle, D., Oppenheimer, D., & Stephens, D. W. (2012). What can evolution do for us?. In *Evolution and the mechanisms of decision making* (Vol. 11, pp. 97–126).

Hebb, D. O. (2005). *The organization of behavior: A neuropsychological theory*. Psychology Press.

Hemmatian, B., & Sloman, S. A. (2020). Two systems for thinking with a community: Outsourcing versus collaboration. In S. Elqayam, I. Douven, J. S. B. T. Evans, & N. Cruz (Eds.), *Logic and uncertainty in the human mind: A tribute to David Over*. New York: Routledge.

Hipólito, I., & Kirchhoff, M. (2023). Breaking boundaries: The Bayesian Brain Hypothesis for perception and prediction. *Consciousness and Cognition, 111*, Article 103510.

Ho, M. K., MacGlashan, J., Littman, M. L., & Cushman, F. (2017). Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition, 167*, 91–106.

Jakob, A. M., & Gershman, S. J. (2023). Rate-distortion theory of neural coding and its implications for working memory. *Elife, 12*, Article e79450.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–292.

Keren, G., & Wu, G. (2015). A bird's-eye view of the history of judgment and decision making. In *The Wiley Blackwell handbook of judgment and decision making* (Vol. 2, pp. 1–39).
Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press.
Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., & Cushman, F. (2020). The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences, 117*(42), 26158–26169.
Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences, 43*, e1.
Lucas, C. (2005). Evolving an integral ecology of mind. *Cortex, 41*(5), 709–725.
Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.
Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron, 36*(2), 265–284.
Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*(4), 979.
Oxford University Press. (2009). *A Dictionary of Psychology* (3rd ed.) Retrieved from https://www.oxfordreference.com/display/10.1093/oi/authority.20110803100539657.
Oxoby, R. J. (2009). The effect of incentive structure on heuristic decision making: The proportion heuristic 1. *Journal of Applied Social Psychology, 39*(1), 120–133.
Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534
Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. Oxford: Oxford University Press.
Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9*(7), 545–556.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.
Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281.
Schwartenbeck, P., FitzGerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 710*(4), 1–5.
Sen, A. (1999). *Commodities and capabilities*. Oxford University Press.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.
Skinner, B. F. (2002). *Beyond freedom and dignity*. Hackett Publishing.
Smith, R., Varshney, L. R., Nagayama, S., Kazama, M., Kitagawa, T., & Ishikawa, Y. (2022). A computational neuroscience perspective on subjective wellbeing within the active inference framework. *International Journal of Wellbeing, 12*(4).
Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. *Topics in Cognitive Science, 00*, 1–28.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189.
Tomov, M. S., Truong, V. Q., Hundia, R. A., & Gershman, S. J. (2020). Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nature Communications, 11*(1), 2371.
Varshney, L. R. (2019a). Mathematical limit theorems for computational creativity. *IBM Journal of Research and Development, 63*(1), 2-1.
Varshney, L. R. (2019b). Must surprise trump information? *IEEE Technology and Society Magazine, 38*(1), 81–87.
Varshney, L. R., Kusuma, J., & Goyal, V. K. (2016). On palimpsests in neural memory: An information theory viewpoint. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications, 2*(2), 143–153.
Von Der Malsburg, C. (1994). The correlation theory of brain function. In *Models of neural networks: Temporal aspects of coding and information processing in biological systems* (pp. 95–119). New York, NY: Springer New York.
Witvliet, D., Mulcahy, B., Mitchell, J. K., Meirovitch, Y., Berger, D. R., Wu, Y., Liu, Y., Koh, W. X., Parvathala, R., Holmyard, D., Schalek, R. L., Shavit, N., Chisholm, A. D., Lichtman, J. W., Samuel, A. D. T., & Zhen, M. (2021). Connectomes across development reveal principles of brain maturation. *Nature, 596*(7871), 257–261.