Title: Reasoning goals and representational decisions in computational cognitive neuroscience: a case study from bounded evidence accumulation

Authors: Ari Khoudary[1-3], Megan A. K. Peters[1-3*], Aaron M. Bornstein[1-3*]

Affiliations:

1. Department of Cognitive Sciences, University of California, Irvine
2. Center for Theoretical Behavioral Sciences, University of California, Irvine
3. Center for the Neurobiology of Learning and Memory, University of California, Irvine

*These authors contributed equally.

Author Contributions:

Ari Khoudary: conceptualization; formal analysis (lead); investigation (lead); methodology; software; visualization; writing – original draft (lead); writing – review & editing (lead). **Megan A. K. Peters**: conceptualization, funding acquisition (equal); methodology, supervision (equal); writing – original draft; writing – review & editing. **Aaron M. Bornstein**: conceptualization; funding acquisition (equal); methodology; supervision (equal); writing – original draft; writing – review & editing

Conflict of interest: The authors have no conflicts of interest to declare.

Data availability: Results from reported simulations as well as all of the code implementing the reported model are available at https://github.com/ari-khoudary/time-varying-precision-weighted-msddm.

Acknowledgements: For several discussions and comments on an early draft, we thank Ainsley May. For helpful comments on later drafts we thank Kevin O'Neill and Dale Zhou. For nurturing nascent philosophical curiosities about formal modeling, A.K. thanks graduate seminars taught by Jeffrey Rouder, Joachim Vandekerckhove, Michael D. Lee, and Mimi Liljeholm. This research was supported by NIMH T32 MH119049 to A.K.

Abstract

Computational cognitive models are powerful tools for enhancing the quantitative and theoretical rigor of psychology and cognitive neuroscience. It is thus imperative that model users—those who develop models, use existing models, or integrate model-based findings into their own research—understand how these tools work, what limitations they have, and what factors need to be considered when engaging with them. To this end, we have developed **ϕ-kit**: a philosophical toolkit for informed, critical use of computational cognitive models in the behavioral and brain sciences. By compiling insights from philosophy of modeling and feminist philosophy of science, ϕ-kit explicates how the *reasoning goals* of a model builder fundamentally shape every step of the model-based research process. We demonstrate the utility of this philosophically-informed perspective by applying ϕ-kit to one of the most popular and successful family of models in cognitive science: accumulator models of two-alternative speeded decision making. The case study begins by investigating notions of optimality both as a general concept and as formalized in accumulator models. Then, we show how different commitments to model optimality (i.e., different reasoning goals) led to development of two “competing” standard forms of the diffusion model and offer a principled heuristic for deciding which form of the model to use. Finally, we demonstrate how insights from ϕ-kit, and the philosophical method more broadly, inspired the development of a novel theory about integration of expectations into decisions that synthesizes previously disparate findings into a common conceptual framework, thus concretely demonstrating how philosophy can advance neuroscientific practice.

Keywords: model specification, optimality, decision making, idealization, abstraction, formal analysis

Epigraph - in memoriam

> There is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination.
>
> —Daniel Dennett, *Darwin's Dangerous Idea*, 1995

Introduction

Cognitive neuroscience and computational cognitive modeling both aim to advance understanding of human cognition by linking behavioral measurements to latent factors hypothesized to be causally relevant in generating that behavior. To do this, it is common to create a controlled experimental environment where researchers measure how behavior changes as a function of manipulating stimuli or task demands. Cognitive neuroscience then correlates task-induced changes in behavior with different measures of neural activity (e.g., event-related potentials measured by EEG or blood-oxygen-level-dependent signals measured by fMRI) to identify latent factors contributing to the behavioral effects. Cognitive modeling, on the other hand, specifies a **model structure**—a set of mathematical objects and operations—and uses behavior to estimate (1) how well that model structure captures task-induced variability in behavior and (2) what values different model components must take on in order to maximize that model's fit to behavior. Much has been written about the reciprocal, complementary relationship between cognitive neuroscience and computational cognitive modeling (Cohen et al., 2017; Forstmann et al., 2011; Forstmann & Wagenmakers, 2015; Kelly et al., 2021; Palmeri et al., 2017; Turner et al., 2017). The burgeoning literature on "model-based" cognitive neuroscience research largely emphasizes the inferential advantages conferred by integrating these approaches—for example, using neural measurements to constrain parameter values or using fitted parameter values to guide interpretation of neural measurements (Forstmann & Turner, 2024; Gluth & Meiran, 2019; Love, 2020; O’Connell & Kelly, 2021; O’Doherty et al., 2007)—relative to using them in isolation (as in traditional cognitive neuroscience or mathematical psychology).

In this article, we contribute to the growing literature on model-based cognitive neuroscience by offering some philosophical perspectives on computational cognitive modeling in cognitive neuroscience. Because models are so powerful and flexible, it is crucial that **model users**—researchers who develop models, use existing models in their own research, or use published model-based research to interpret their own results—understand how these tools work and what factors to consider when engaging with them. To this end, we have developed *ϕ-kit*: a philosophical toolkit crafted to promote informed, critical use of computational cognitive models in the behavioral and brain sciences. *ϕ-kit* addresses basic ontological questions about what a model is and how models relate to their **targets** (empirical phenomena of interest) by summarizing some consensus views from philosophy of modeling. We further explicate epistemological considerations that follow from these consensus views, identifying different factors that shape and/or constrain what models can tell us about targets. Throughout, *ϕ-kit* emphasizes the role of **reasoning goals**—how a user wishes to reason about their target when using a model to do so—in assessing model performance and interpreting model-based findings. This position follows logically from the consensus view on models’ ontological status. We also hope that emphasizing the role of reasoning goals will remind readers that model-based research involves a large degree of decision-making on the part of the researchers, and offer some guidelines for thinking about what factors shape particular decisions other researchers have made.

We further demonstrate the utility of *ϕ-kit* by applying it to one of the most popular and successful family of models in psychology and neuroscience: accumulator models of decision making. We present the results of our philosophical explication as a case study comprised of several lessons. The first two lessons examine the notion of optimality: what it means to claim that a system is performing optimally, and how this notion is operationalized in accumulator models. The next three lessons use an open debate in the literature to demonstrate how reasoning goals shape the representational decisions researchers make when building and evaluating models. Specifically, we investigate the topic of decision boundaries that collapse over time and offer a principled heuristic for deciding whether to incorporate this feature into one's model. The case study concludes with two lessons demonstrating how insights from *ϕ-kit* led to the development of a novel theory, which we implement as an accumulator model. This theory—directly motivated by considerations about the goal-dependence of empirical findings—offers a unifying explanation for previously disparate findings about effects of expectations on decisions change as a function of time. Altogether, the case study demonstrates how both the outputs and methods of philosophical analysis are indispensable to the practice of computational cognitive neuroscience.

***ϕ-kit*: a philosophical toolkit for computational cognitive modeling**

The consensus view in philosophy of modeling posits that models are representational tools for reasoning whose performance is assessed by *adequacy for purpose* and not verisimilitude (true or accurate representation of the natural world) (Frigg & Hartmann, 2020; Parker, 2020). This is precisely the position captured in the famous aphorism "all models are wrong, but some are useful" (Box and Draper, 1987, p. 424). The adequacy-for-purpose view of models is a pragmatic solution to the fact that models are not *truth-evaluable* entities: they cannot be judged as true or false (Wimsatt, 1987). Models are not truth-evaluable because they are representations of targets (i.e., objects or phenomena in the natural world), and thus always[1] involve some degree of falsity with respect to the target (Parker, 2020; Winsberg & Harvard, 2024). While this might seem like a reason *not* to use models, it is precisely their representational nature that makes models useful. A representation that is perfectly verisimilar to a target is just a clone of that target, which is not always a helpful tool for scientific reasoning about the target.[2] When verisimilitude is cast aside as a guiding principle, the goal of modeling becomes identifying representations that are *useful for reasoning* about targets. Whether a particular representation is useful for reasoning further depends on one's *inferential* or *reasoning goal* in using the model (e.g., predicting future behavior, identifying loci for causal intervention, generating different types of explanations). In order to identify models that are useful for a particular reasoning goal, one must understand (1) *how* representations permit reasoning about the targets they represent and (2) *what* kinds of reasoning different representations permit of their targets. We address these in turn.

Precisely how representations figure in scientific reasoning is an area of open research in philosophy of modeling. For our purposes, it suffices to say that representations permit reasoning about targets by providing scientists with a simplified, often idealized version of the target that they can interact with *as if* it were the actual target. By omitting some properties of

[1] One exception is scale models of physical structures (e.g., a three-dimensional replica of a building). These models have the ability to accurately represent each component of the target, just at a smaller physical scale. We don't address them here because scale models are of a different type (i.e., concrete or material) than those that are the focus of this toolkit (formal/mathematical), and are more commonly used in engineering than neuroscience.

[2] This sentiment is captured by another famous aphorism "A map is not the territory it represents" (Korzybski, 1933). Models are commonly analogized to spatial maps because maps provide an intuitive concept of the utility of representation. Winsberg and Harvard (2024) use the example of a subway map, which only needs to represent the *order* of stops and not necessarily the distance between them or details about the path the train takes to get from one stop to another. Omitting these details from the representation makes the subway map quickly and intuitively useful, but only for purposes of navigating the subway system; navigating the actual city will require a more detailed (or less abstract) map.

the target and idealizing others, models create representations of targets that "selectively attend" to components of the target that the user wishes to reason about (Portides, 2021). The behavior of this hypothetical target can then be compared to that of the actual target to see how adequately that representation captures properties of the target about which the modeler aims to reason. In formal modeling, a common practice is to construct several possible models of a target process and use quantitative metrics (e.g., Bayesian information criterion; Schwarz, 1978) to identify the model that most succinctly accounts for the observed data.

Different types of representations permit different types of reasoning about targets. Some representations are physical (e.g., whiteboard drawings, scale replicas, awake behaving rodents), and others are abstract (e.g., mathematical equations). All of these representations can function as models, and surveying the types of reasoning each permits is far beyond the scope of this paper. Interested readers can consult Winsberg and Harvard (2024) and Frigg and Hartmann (2020) for helpful reviews on this topic. The present article is concerned with the types of reasoning facilitated by *formal* (and/or mathematical) models. These models are commonly implemented via code or software, making them *computational* as well. The distinction between a formal model and its implementation is important but beyond the scope of this paper; we refer readers interested in these considerations to Guest and Martin (2021) and Cooper and Guest (2014). Finally, we restrict our focus in the toolkit to *formal cognitive models* of the kind historically developed in the field of mathematical psychology (Batchelder, 2010). Models of this type are commonly thought to instantiate formal theories about the latent cognitive structures and processes driving behavior (Grahek et al., 2021; Guest & Martin, 2021; Navarro, 2019; Press et al., 2022; van Rooij & Baggio, 2021).

A consensus view in the meta-scientific literature is that formal models enhance the rigor of behavioral sciences by allowing modelers to reason about their targets using the quantitative and conceptual precision of formal languages (Devezer et al., 2021; Grahek et al., 2021; Guest & Martin, 2021; Lee et al., 2019; Robinaugh et al., 2021; Smaldino, 2020; van Rooij & Blokpoel, 2020). Formal languages, by design, do not inherently refer to objects or properties in the natural world—this is precisely what makes formal models *abstract*. In order to reason about a target with a formal model, a model user (i.e., a human who builds or interprets a model) has to ***decide*** how they will map components of the model onto components of the target: for example, deciding that the coupled differential equations in the drift diffusion model represent the evolution of a belief rather than the flow of ions into a neuron. This mapping is known as the model's *construal*, whereas the mathematical objects and equations comprising the formal representation are called the model's *structure*[3] (Andrews, 2021). In the case study, we show how recognizing the difference between model structure and construal can offer clarity about when one ought to make particular representational decisions.

Although structure and construal are separable properties of a model, the model's structure places important constraints on the space of plausible construals. For example, one cannot reasonably construe a model structure with no temporal dynamics as representing the evolution of some quantity over time—the model simply does not have the representational resources to support this kind of reasoning about the target. The solutions in this case are either to change one's reasoning goals so that they no longer include dynamics of the target's behavior, or modify the model structure so that it can represent how the relevant features might change over time. This simple example highlights the necessity of aligning mathematical representation with one's *reasoning goals* when engaging with a model, which in turn requires

[3] Although see Levenstein et al. (2023), who embed construal into their definition of a model!

that users have a clear idea of what exactly their reasoning goals are when they engage in model building, development, and testing.

With some notable exceptions (Blohm et al., 2020; Kording et al., 2018), reasoning goals have been relatively underappreciated in computational neuroscience. Discussions about the epistemic utility of different models have largely focused on the *type of model* instead of the *type of reasoning* it permits. An example of this can be found in Dayan and Abbott (2005), who distinguish among *descriptive, mechanistic,* and *interpretive* (more commonly called *normative*) models in neuroscience. On their taxonomy, descriptive models accurately and compactly summarize large amounts of data, mechanistic models describe how functional and anatomical properties of the nervous system produce different behaviors, and normative models use formal principles to investigate why nervous systems operate how they do. These “types of models” can be straightforwardly translated into reasoning goals: describing effects of experimental manipulations on the target’s behavior, identifying mechanisms that underlie those changes in behavior, and investigating why the target responded in the specific way that it did (instead of some other way). It is important to note, however, that the variety of possible reasoning goals is much broader than indicated by this taxonomy (see Kording et al., 2018 for an empirical survey[4] and Wimsatt’s (1987, pp. 7–8) “Twelve things to do with false models”). Two of these broader goals, for example, are (1) exploration of unconceived alternative explanations for empirical phenomena (Stanford, 2006; Wimsatt, 1987) and (2) identifying

[4] It is worth noting that—because their paper focuses on “goals” broadly construed, rather than “reasoning goals” specifically—some goals in the Kording et al. (2018) are better understood as *epistemic values* within the context on this article. For example, they define compactness, analytic tractability, interpretability, and beauty as *goals* of modelers, whereas we consider them *values* of the modelers. We take these ideals to be values instead of reasoning goals because they apply to a model’s *structure* rather than its intended construal.

endpoints in the space of possible explanations, where the “true” explanation that lies between the endpoints is too complex to specify or analyze (Wimsatt, 1987).

After deciding on a reasoning goal, model-builders must undertake a series of *representational decisions*—deciding *what* to represent in the model and *how* to represent it (Harvard & Winsberg, 2022; Winsberg & Harvard, 2024)—in order to create a model that is adequate for their purpose. Answers to the first question determine the *level of abstraction* at which the model operates. There are two relevant senses of abstraction at play here: one pertaining to the level at which the research question is posed (e.g., asking about single cells or whole-brain activity), and one pertaining to the level of detail or specificity in the mathematical representation (e.g., whether neurobiological details will be rendered). Importantly, questions at the same level of physical abstraction can be addressed with models at different levels of mathematical abstraction. Memory retrieval, for example, can be successfully modeled using relatively abstract models based in signal detection theory (Schurgin et al., 2020) or more detailed models based in theoretical principles specific to the process of memory retrieval (Howard & Kahana, 2002). Again, the “correct” model here is simply the one that better aligns with one’s reasoning goals—in this case, whether one is interested in reasoning about how memory retrieval is instantiated as a domain-general decision process or as a unique cognitive phenomenon that requires additional theoretical considerations.

Considering *how* to represent components of the target mathematically is at the heart of model building. One of the first steps here is deciding on (or developing) a *formal framework* within which the model will be built. Formal frameworks consist of axioms/postulates (i.e., statements accepted as true), mathematical objects, and rules governing interactions between and transformations of the mathematical objects. Formal frameworks can be thought of as providing a “grammar” for formal theory development.[5] Any representation built using the tools of a particular formal framework must abide by the axioms of that framework in order for the model to be valid.[6] But because axioms function to promote mathematical expressivity (the ability to write down particular ideas in math) rather than verisimilitude, model builders are often forced to construct *idealized*—rather than veridical—representations.

Philosophy of modeling distinguishes between two types of idealization which we will call *omissive* and *disortive*.[7] Sometimes referred to just as the process of abstraction (Jones, 2005), omissive idealizations are those that omit particular features of the target in order to better align the model with the formal framework (e.g., modeling neurons as a sigmoid function in neural networks when we know that their functions are more complex than a sigmoid transform). Distortive idealizations, on the other hand, deliberately *misrepresent* properties of the target in order to make it more mathematically tractable (e.g., assuming observers have full or perfect knowledge of their environment when we know that these conditions hardly ever obtain in the real world). Most models include both types of idealizations. A concrete example can be found by foreshadowing our argument in the case study. The formal framework underpinning a prominent family of perceptual decision-making models requires particular distortive idealizations that logically imply particular omissive idealizations. These omissive idealizations, in turn, largely preclude any formal consideration of the role that uncertainty about one’s expectations might play in expectation-guided perceptual decisions.

[5] Some common formal frameworks in psychology and neuroscience include signal detection theory, sequential analysis, information theory, Bayesian inference, and Markov decision processes. Guest & Martin (2021) offer a helpful analysis of how formal frameworks relate to theories and models in cognitive psychology.
[6] A familiar example here comes from the assumption of normally-distributed residuals at play in general linear regression models. If one’s data do *not* exhibit normally-distributed residuals, then the results of the general linear regression are not guaranteed to be valid.
[7] These types of idealization have historically been called “Aristotelian” and “Galilean”, respectively (Cartwright, 1983; Winsberg & Harvard, 2024), in honor of the scientists whose research heavily involved use of the eponymous idealization.

Thus, we can understand models as the end-result of a series of *representational decisions* made by model builders (Blohm et al., 2020; Harvard & Winsberg, 2022). After choosing an appropriately specific model target, modelers have to decide *what* they want to represent of their target and *how* they wish to represent those properties in the model (Winsberg & Harvard, 2024). Doing so effectively requires that modelers specify their *reasoning goals* in building the model. On the adequacy-for-purpose view, reasoning goals are at the heart of the model evaluation process. Harvard and Winsberg (2022) further make the point that a researcher's reasoning goals reflect their *epistemic values*: for example, what they think the science aims to achieve, how important particular methodological limitations are for conducting particular inferences, and what is required for a good explanation given these considerations. Adherence to a particular theory and/or model structure can also function as an epistemic value that biases researchers' evaluations of alternative models; we provide examples of this in the case study. This phenomenon is commonly referred to as the *theory- or value-ladenness* of scientific reasoning, which captures the facts that (1) scientific knowledge is the product of human thought, and (2) human thought is fundamentally shaped by idiosyncrasies of values and experiences regardless of how objectively any one individual aims to reason (Boyd & Bogen, 2021). A common solution to the problem of value-ladenness is to integrate a diversity of perspectives into the process of scientific inquiry (Longino, 1998), which has been shown to mitigate biases introduced by individual scientists (Wu, 2023). We take this to be true for modeling goals as well.

Altogether, the considerations above indicate that while there are some objective heuristics that model builders can use to quantify the utility of their decisions (e.g., goodness of fit, out-of-sample predictions, convergence criteria), the bigger question of which mathematical representation is *adequate for purpose* fundamentally reduces to one's reasoning goals in building the model. As we will show in the case study, disagreements and/or confusion about which model is "correct" often reduce to unrecognized differences in reasoning goals among researchers rather than objective properties of the models themselves.

Introducing our target: accumulator models of speeded decision making

We now turn to getting acquainted with the target of our case study: accumulator models of decision making. This is a family of models that represent decision-making as a process of evidence accumulation, or adding up information over time. Models in this family are conceptually united by the *sequential sampling framework* in psychology and neuroscience (Forstmann et al., 2016; Gold & Shadlen, 2007; Ratcliff et al., 2016; Shadlen & Shohamy, 2016), which itself draws on the framework of *sequential analysis* in statistics (Barnard, 1946; Wald, 1945).[8] The conceptual framework of sequential sampling posits that humans and other animals make decisions by continuously sampling information from an evidence source, extracting and integrating decision-relevant information over time, and committing to one of the options once the accumulated evidence surpasses a threshold value. The formal framework of sequential analysis permits specifying normative solutions to the accumulation process, and thus can be used to build models that optimize the tradeoff between decision accuracy and deliberation time (i.e., the *speed-accuracy tradeoff*). Importantly, however, not all models in this family are normative. A major strength of the sequential sampling framework is the generality and flexibility of its conceptual entities (e.g., "evidence"), both of which permit users to specify

[8] An interesting and relevant bit of history is that two different goals led to independent developments of the sequential analysis framework during World War II. One goal was enhancing efficiency of industrial output (Barnard, 1946; Wald, 1945), and another was cryptanalysis to decode enciphered German messages. This latter method was derived by Alan Turing, and the relationship of Turing's framework to sequential sampling is detailed quite nicely in (Gold & Shadlen, 2002).

a wide range of models that can be construed at multiple levels of physical abstraction. For our purposes, we can think of each model in the family as a formal theory about the the architecture of the decision process (e.g., the shape of the decision thresholds, connections among accumulation units, etc.) and the mechanism(s) whereby evidence is accumulated (e.g., continuous vs. discrete-time integration, loss of early information, corruption by noise, etc.).

Grounded in a rich history of development within mathematical psychology (Busemeyer & Townsend, 1993; Laming, 1968; Link & Heath, 1975; Stone, 1960; Vickers, 1970), accumulator models currently function as the default approach for formal modeling of reaction times during decision making. This precedent was facilitated by the development of an experimental paradigm that permits straightforward mapping of stimulus features onto components of the models: random dot motion discrimination (Britten et al., 1992; Gold & Shadlen, 2001). In this task, observers are presented with a display of stochastically moving visual elements (usually dots), some proportion of which consistently move from one direction to another (e.g. left to right). On each trial, observers report which direction of motion appeared on the display, a judgment whose difficulty scales with the proportion of consistently moving dots (i.e., the *coherence* of the stimulus; Palmer et al., 2005). This combination of a precise and flexible psychophysical paradigm with a simple modeling framework created fertile grounds for robust theory-guided research. Indeed, a series of careful early experiments on non-human primates performing the random dot motion task revealed that both single-unit and population level recordings from distinct cortical regions exhibit activity that strongly resembles and correlates with dissociable conceptual components of accumulator models (Britten et al., 1996; Gold & Shadlen, 2001; Roitman & Shadlen, 2002; Shadlen & Newsome, 2001). Since these early findings, evidence accumulation has been used to link behavior with population-level neural responses in rodents (Brunton et al., 2013; Hanks et al., 2015), as well as intracranial recordings, scalp oscillations, and BOLD signals in humans (Krueger et al., 2017; Polanía et al., 2014; Weber et al., 2024).

While powerful, this wide range of empirical success is also a primary source of confusion or misunderstanding about the claims a researcher aims to make when using this kind of model, as well as what kind of data are relevant for testing their claims. These challenges are exemplified by the vast literature and open debates about the appropriate mathematical structure of models in this family (Donkin et al., 2011; Hawkins et al., 2015; Khodadadi & Townsend, 2015; van Ravenzwaaij et al., 2012). A substantive barrier to progress is the rather high degree of *mimicry*—identical behavioral predictions from theoretically different structures—among the models (Bogacz et al., 2006; Donkin et al., 2011; Khodadadi & Townsend, 2015; O'Connell et al., 2018). A related challenge comes from the fact that conceptually different structures can also exhibit *formal equivalence*, meaning not only that they make identical behavioral predictions, but also that they are *mathematically indistinguishable* from each other. One example of this is the incorporation of a quantity that changes in magnitude over the course of a single decision, sometimes called an "urgency signal" (Churchland et al., 2008; Murphy et al., 2016; Thura et al., 2012). Conceptually, this quantity can be incorporated either as a decrease in the decision boundary (such that it adjusts the decision rule) or a gain on the decision variable (such that it modifies the internal representation of evidence); but mathematically, it is impossible to specify model structures that dissociate these possibilities.

For all of these reasons, accumulator models offer a strong example of (1) how models can be used to investigate specific hypotheses about latent processes driving behavior, (2) how there can be disagreement about how scientists think the models should do that, and (3) how scientists' positions on (2) influence the representational decisions they make when building and assessing competing models in this family. And, from a practical perspective, their popularity and the dearth of critical perspectives in the model-based cognitive neuroscience literature make them prime targets for philosophical explication. Toward this end, our case study uses *ϕ-kit* to gain some traction on a long-standing stalemate about diffusion-to-bound accumulator models: whether the decision rule remains fixed across trials while other parameters can vary trial-by-trial (Ratcliff & McKoon, 2008), or if the decision rule changes dynamically within a trial while other parameters remain constant across trials (Drugowitsch et al., 2012).

BOX 1:

Among the dozens of existing sequential sampling models, one specification—the *diffusion decision model* (or *drift diffusion model*; both abbreviated DDM)—has become the standard form (Ratcliff et al., 2016). The model is specified by four free parameters: starting point (red dot in Figure 2A), non-decision time (yellow line in Figure 2A), boundary/threshold separation (green lines in Figures 2), and drift rate (blue arrows in Figure 2). The model also relies on the concepts of momentary evidence (stimulus/environmental information sampled over the course of a decision), diffusion (the amount of noise in an observer's encoding of momentary evidence), and a decision variable (the observer's internal representation of accumulated information; black and gray lines in Figure 2). The starting point variable defines the point along the y-axis where the decision variable begins its process of evidence accumulation, and thus is conventionally interpreted as quantifying the bias an observer has toward one of the two choice options. The non-decision time variable quantifies the time needed for an agent to perform processes unrelated to evidence accumulation, like the time needed to extract decisional evidence from momentary evidence and the time needed to prepare a motor action after internally committing to a choice (i.e., the decision variable hitting one of the thresholds). The threshold separation variable defines the quantity of accumulated evidence required for an observer to commit to a choice, and thus is commonly interpreted as reflecting the observer's response caution or decision policy under different speed-accuracy regimes. Finally, the drift rate captures the slope of the decision variable, and is often interpreted as reflecting the "quality" (i.e., signal-to-noise ratio) of momentary evidence or, more specifically, quantifying how much each piece of momentary evidence contributes to the ultimate decision.

Although the DDM was initially developed using static values for the starting point, drift rate, and non-decision time (Ratcliff, 1978), the contemporary standard form of the model incorporates trial-level variability in each of these parameters. Starting point and non-decision times are typically drawn from a uniform distribution, whereas drift rates are drawn from a normal distribution (Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). This variability allows the model to better capture differences in RT distributions between correct and incorrect trials. Drift rate variability captures slow errors, starting point variability captures fast errors, and non-decision time variability improves model fits when there is a large degree of across-condition variability in the 0.1 quartile of the RT distribution (Ratcliff & McKoon, 2008; Ratcliff & Tuerlinckx, 2002). Following common parlance, we use "original DDM" to refer to the model without trial variability, and "extended DDM" to refer to the model with trial variability. Because the original DDM fails to capture important empirical properties of RT distributions, the extended DDM has become the current standard form. A popular alternative to the extended DDM is a version of the model without trial-variability in the relevant parameters, but with a boundary separation value that decreases over the course of a decision (Figure 2B). This "collapsing bound" form of the model provides many of the same empirical advantages as the extended DDM, but is specified in a way that posits a different underlying mechanism of the speed-accuracy tradeoff: namely, that observers become more liberal in their decision threshold as a function of elapsed decision time (Drugowitsch et al., 2012; Frazier & Yu, 2007).

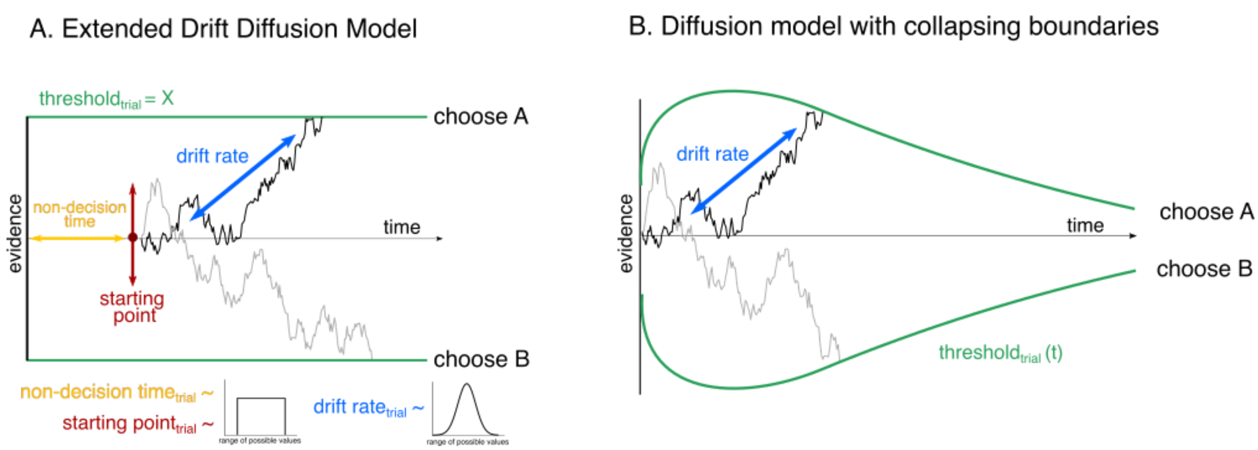

FIGURE 2: Graphical depictions of two standard forms of a diffusion model of speeded two-alternative choice. Both models assume that observers sequentially sample information, accumulate the difference in evidence between the two choice options, and commit to a decision once the accumulated evidence reaches a critical value defined by the decision boundary/threshold. Standard interpretations of the key variables are provided in the main text. (A) The extended drift diffusion model. This form of the model features parameters whose values vary trial-by-trial but remain fixed within the course of a single trial. Starting point (yellow) and non-decision time (red) values are drawn from uniform distributions and drift rates (blue) are drawn from normal distributions. Threshold values (green) are constant within a trial,

and can be allowed to vary as a function of experimental condition and/or participant. (B) The diffusion model with collapsing decision boundaries. This form of the model features parameters whose values are effectively fixed trial-by-trial and decision boundaries whose values decrease ("collapse") over the course of a single trial. Drift rate and, in some models, starting point are permitted to vary as a function of experimental condition but are assumed to have fixed effects across trials within an experimental condition.

Applying the toolkit to our target

Now that we are also familiar with key components of accumulator models, we can begin applying the toolkit to our target. Throughout seven "lessons", we investigate how different reasoning goals—informed by different epistemic values among researchers—shape the representational decisions those researchers made when building and evaluating competing models of behavior. Lessons 0 and 1 explicate the concept of optimality and how it is formalized in accumulator models, respectively. Lesson 2 highlights an informative mismatch between the stated goals behind the original DDM and its contemporary status as the optimal two-alternative decision making process. Lesson 3 then examines how different researchers modified the original DDM after deeming it *in*adequate for their purpose, thus leading to the stalemate about the "correct" form. Next, Lesson 4 demonstrates how the procedures one uses to compare these two models are also shaped by the reasoning goals and epistemic values of the researchers performing the model comparison. An intermediate summary integrates insights developed throughout the lessons to provide a principled heuristic for deciding which form of the diffusion model is suitable for one's reasoning goals. Lesson 5 then advances an argument that extant approaches for measuring and modeling expectations in perceptual decision making are critically limited by choices stemming from idealizations. Finally, Lesson 6 presents a novel model we developed to reason about principled decision processes in less-idealized environments and shows that this model synthesizes two disparate observations of time-varying effects of expectations on perceptual decisions.

Lesson 0: explicating optimality

Reasoning goals in the scientific study of decision making—and perceptual decision making in particular—commonly invoke the notion of optimality. This term has a rather intuitive meaning in everyday language: performing as best as possible. Scientific usage of this term, however, requires appeal to a number of tacit assumptions about *what* the system ought to be doing and *how* it ought to be doing that. Unsurprisingly, then, invoking optimality in one's research question or explanation can lead to talking-past among different research groups (Rahnev & Denison, 2018) and/or conclusions that are not as strongly justified as the researchers might believe (Eberhardt & Danks, 2011). A brief explication about the concept of optimality can offer some salve for these issues; or at the very least, it will provide useful foundation for issues considered in the rest of the case study.

Optimality is commonly used to ask and answer normative questions about one's target, i.e., investigating *why* it behaves in *this particular way* (instead of some other equally plausible way). Recall that formal frameworks (e.g., sequential analysis) permit specifying normative solutions to formally-specified problems (e.g., deciding between two possible outcomes by accumulating evidence over time). A formal specification of the problem a system is hypothesized to be solving is commonly called the *objective* function of that system.[9] A normative solution is one that *optimally solves the problem,* i.e., identifies the mathematical limit of the system's performance with respect to the objective function and the environment.

[9] The objective function is also commonly called a *cost* or *loss* function.

Normative models thus answer *why* questions by appeal to the objective function: the system behaves this way (and not some other way) because this behavior optimizes *this* objective function (and not some other objective function).

Lesson 1: defining optimality in accumulator models

Thus, the choice of objective function is a core representational decision in the process of normative modeling. Because it defines the criterion of optimality against which the target system will be compared, it is imperative to align the specification of objective function with the type and scope of normative reasoning one wishes to perform with respect to the target. Effectively all theoretical work on accumulator models assumes that decision makers aim to optimize a *speed-accuracy tradeoff*. While this is a reasonable choice, contemporary work rarely, if ever, offers explicit justification for choosing this objective function (e.g., Drugowitsch, Moreno-Bote et al., 2012; Tajima et al., 2019). Absent explicit justification or discussions of alternative notions of optimality, the literature risks losing sight of the fact that optimizing speed-accuracy tradeoffs is not a necessary feature of accumulator models. We highlight alternative approaches to optimality analyses in Lesson 4. But first, this lesson introduces the two formalizations of speed-accuracy optimality at play in accumulator models.

The first formalization of a speed-accuracy tradeoff comes from the sequential probability ratio test (SPRT) developed independently by Barnard (1946) and Wald (1945). In SPRT, the decision variable is a running estimate of the log-likelihood ratio of one choice option relative to another, and the threshold values are fixed across a decision. Shortly after its introduction, SPRT was mathematically proven to minimize *Bayes Risk*: a weighted, linear sum of decision time and error rate (Bogacz et al., 2006; Wald & Wolfowitz, 1948). Thus, the speed-accuracy tradeoff formalized by SPRT is defined as minimizing the *average* time needed to produce responses at a fixed level of accuracy. Crucially, this proof assumes that the "difficulty" of the decision (e.g., the signal-to-noise ratio in the stimulus) is identical for all trials included in the estimate of average reaction time—i.e., that the decision environment is *homogenous* (Moran, 2015). When the decision environment is heterogeneous (i.e., different difficulty on each trial), SPRT no longer minimizes Bayes Risk and is therefore no longer optimal.

Rather than minimizing Bayes Risk, it is currently more common for modelers to assume that decision makers are *maximizing* their *reward rate*—average reward per unit of time (Gold & Shadlen, 2001, 2002). If an experiment does not deliver performance-based rewards to decision-makers, then correct and incorrect answers are modeled as rewards and punishments, respectively. Importantly, the formalization of time in this criterion sums up *all* the temporal components of the decision environment: decision time, non-decision time, inter-trial interval, and potential time penalty for delays. This more granular representation of time gives reward rate optimality a number of advantages relative to Bayes Risk. First, and most saliently, the validity of reward rate as a criterion of optimality does not require that decision difficulty remains fixed across trials. Next, its representation of time permits theoretically and empirically investigating how different environmental dynamics shape choice behavior. Further, by invoking the notion of reward, it aligns accumulator models more closely with other formal models of decision making (e.g., expected utility theory) which enhances the prospects for inter-theoretic model building. Additionally, it has been argued to represent a more ecologically valid concept of optimality for human and animal decision makers (Balci et al., 2011; Moran, 2015). Finally, it has the practical advantage of being parameter-free: assuming that decision makers optimize reward rate does not require specifying how much weight a particular agent places on accuracy vs. speed (Bogacz et al., 2006).

Intermediate summary: lessons on optimality

The previous two lessons had several aims. Lesson 0 briefly explicated the notion of optimality to ground future discussions in the case study and invite critical reflection on its use as an explanatory concept. Lesson 1 then introduced two standard formalizations of optimality in accumulator models to provide concrete examples of considerations in Lesson 0. Together, the lessons aim to highlight that the specific conception of optimality in any modeling framework is a *choice* submittable to the same interrogations we perform on other experimental and/or modeling choices, despite it *prima facie* seeming like a fixed property of a particular model. Additionally, we aimed to demonstrate that a concept as simple as optimizing the speed-accuracy tradeoff can be formally expressed in a number of different ways, each of which yield subtly different claims about exactly what the target system ought to be doing. As we will show in the rest of the case study, careful consideration about how one's conceptual reasoning abilities are constrained by representational decisions required by formalization and specification is imperative for critical and informed model-based neuroscience research.

Lesson 2: incidental optimality of the DDM

The paper introducing the original DDM (Ratcliff, 1978) is titled *A theory of memory retrieval.* On the first page, Ratcliff (1978; p. 59) states his reasoning goal: "The theory presented in the present article is concerned with providing an account of processes underlying retrieval from memory." He goes on to write, "The theory presented here has been designed to deal with all aspects of the data in each paradigm (accuracy, response latency, and latency distributions) and has been designed to apply over a range of paradigms" (Ratcliff, 1978, p. 60). Thus, it is exceedingly clear that the original goal behind developing the DDM was to specify a *descriptive* model of *memory retrieval*. The words “optimal” and “normative” do not appear anywhere in the text. However, the DDM currently functions as the *normative solution* to *speeded decision making*, particularly of the perceptual variety (Bogacz et al., 2006). When and how did this shift occur?

In a highly influential paper, Bogacz et al. (2006) conducted a systematic mathematical analysis of popular accumulator models, comparing them to each other and to different optimality criteria. They showed that the original DDM is the continuous-time equivalent of SPRT, making it the optimal procedure for minimizing Bayes Risk. The authors also showed that for other criteria of optimality (e.g., reward rate and balancing reward with accuracy), there is a unique value of the DDM threshold that optimizes the decision process according to that criterion. Further, they found that five of the six alternative accumulator models all reduce to the original DDM when their parameters optimize reward rate (Bogacz et al., 2006). These results offer strong evidence for the optimality of the DDM, both in terms of choice behavior and decision dynamics.

Thus, a model that was explicitly developed to provide a task-general description of memory decisions, and later to offer mechanistic insight into the formation of domain-general speeded two-alternative decisions, was also found to provide a normative solution to the speed-accuracy tradeoff in two-alternative decision making. We find this story particularly remarkable because across the sciences, and perhaps especially in the cognitive sciences, it is rarely the case that a mathematical model developed for description also provides a normative solution—normative solutions are typically pursued as ends in their own right. And from an inferential perspective, the optimality results raise substantive implications for targets successfully modeled by the DDM: they optimally accumulate evidence over time.

Lesson 3: The value-ladenness of model specification

In this lesson, we review how different groups of researchers have modified the original DDM after deeming it *in*adequate for purpose. We focus on the development of the two current standard forms of accumulator models—the extended DDM (e.g., Ratcliff and McKoon, 2008) and a diffusion model with collapsing boundaries (e.g., Drugowitsch, Moreno-Bote et al. (2012)—highlighting how stated differences in reasoning goals have resulted in the two standard forms.

As reviewed above, representational choices in the DDM have been grounded in the reasoning goal of providing a unified mathematical description of core empirical properties of two-alternative decisions. In particular, the model has been developed such that it can reproduce differences in the shapes of reaction time distributions for correct and incorrect decisions. This goal explicitly motivated the addition of trial-variability in drift rate and starting point in the extended DDM: “The key feature of the model that allows it to deal with the complexities of error response times is the assumption that parameters of the model (drift rate and starting point) are variable from trial to trial.” (Ratcliff and Rouder, 1998, p. 356). Further, the goal of adding variability in the non-decision term was to enhance the model’s ability to fit values in the 0.1 quantile of RT distributions across conditions in a task (Ratcliff & Tuerlinckx, 2002), which it successfully accomplished. Thus, the current standard form of the DDM—one with trial variability in drift rate, starting point, and non-decision time—reflects a series of representational decisions anchored in the goal of developing a unified, statistically robust mathematical description of the empirical properties of two-alternative choice behavior.

This series of representational decisions can be contrasted against those on display in a highly influential paper by Drugowitsch, Moreno-Bote et al. (2012). The conceptual launching point for these authors was inadequacy of the *assumptions* under which the DDM optimizes the dynamics of two-alternative deliberation. Specifically, the authors modified two key assumptions underlying previous proofs of DDM optimality: homogeneity of decision environments (assumed by Wald & Wolfowitz (1948) for optimality of SPRT) and no cost for accumulating evidence (assumed by Bogacz et al., 2006). Their stated goal in doing so was to "provide a formalism to determine the *optimal* behavior given a total description of the task, the rewards, and the costs" (Drugowitsch, Moreno-Bote et al., 2012, p. 3612; emphasis added). To do this, they specified a heterogeneous environment and a reward rate function that incorporated a small cost for each sample drawn over the course of a decision. Using techniques from dynamic programming (i.e., the Bellman equation), Drugowitsch, Moreno-Bote et al. (2012) found that the optimal solution on these assumptions is to accumulate evidence according to the diffusion process[10] and dynamically adjust the separation of choice thresholds during the course of a decision (Figure 2C). Importantly, this form of the model does not require trial-variability in drift rate, starting point, or non-decision time; the goodness-of-fit advantage provided by variability in these parameters was achieved by the time-varying decision boundary alone.[11]

[10] It's worth noting that when the assumption of constant task difficulty is relaxed (i.e., when the decision maker does not know exactly what the strength of evidence on that trial is), the decision variable can no longer be interpreted as encoding the log-odds of either choice being correct (Drugowitsch, Moreno-Bote et al., 2012, p. 3622). This is a subtle but crucial point for construing diffusion models applied to data acquired in heterogeneous environments (i.e., the majority of contemporary experimental conditions).
[11] For the sample of participants in the task reported in Drugowitsch, Moreno-Bote et al. (2012). The authors did not compare performance against the extended DDM. Subsequent work has shown that collapsing boundaries coupled with trial-variability in drift rate enhances model fits relative to collapsing boundaries without drift variability (Murphy et al., 2016; O'Connell et al., 2018).

Thus, we find ourselves with two model structures of the same target process. The models make highly similar predictions about most aspects of the data but specify qualitatively distinct neurocognitive architectures underlying two-alternative decision making. The theories formalized in these models reflect the reasoning goals of the researchers who built them. In the DDM, which was originally motivated by a reasoning goal to unify empirical phenomena, researchers put the data first and iteratively adjusted the model structure until it maximally explained empirical properties of a broadly-scoped target. Crucially, the subsequent representational changes broke the normative status of the original model: trial-variability in starting point, drift rate, and non-decision time make the DDM no longer equivalent to SPRT, and thus no longer the optimal procedure for minimizing Bayes Risk (Moran, 2015). Contrast this approach with that of Drugowitsch, Moreno-Bote et al. (2012) who explicitly pursued a formal structure that permits normative reasoning about the target system under slightly less idealized assumptions. Their form of the model resulted in model fits on par with those generated by the extended DDM in similar task environments, and also provides representational resources to capture an increasingly-documented neural phenomenon: a time-varying gain on the responses of neurons postulated to encode the decision variable (Churchland et al., 2008).

As scientists, we are likely tempted to ask which of these competing models is true. Our toolkit suggests that this is the wrong question; or at least, that the question is ill-posed. We expand upon this position in the next lesson.

Lesson 4: The value-ladenness of model comparison

One way to conceptualize the contrast drawn in Lesson 3 is between representational decisions based on statistical vs. theoretical considerations. This lesson continues along that theme to conceptually analyze the practice of *model comparison*, a central but often under-discussed step in the process of model-based scientific inference. We begin by reviewing a controversial paper comparing diffusion models with fixed and collapsing bounds (Hawkins et al., 2015), and then contrast its approach to a paper using formal reasoning to compare the utility of different model structures (Moran, 2015).

In response to the growing popularity of collapsing-bound accumulator models, Hawkins et al. (2015) undertook a large-scale quantitative comparison of model performance on a variety of perceptual decision making datasets. The authors took great care to ensure the statistical robustness of their results, utilizing data from different research groups and species, specifying multiple forms of collapsing bounds models, and running several computationally intensive sensitivity analyses. Further, they reported that three different quantitative metrics of model performance—Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and nested likelihood ratio tests—all returned similar results, which was their justification for reporting only the BIC values in the published manuscript. Importantly, BIC is a heuristic metric that aims to approximate Bayes Factors and tends to overpenalize model complexity (when complexity is formalized as the number of parameters in the model[12]). This made it a particularly fraught choice for their first round of analyses, where the collapsing bounds models were all specified in a manner that made them between 1.3 to 2 times as complex as the extended DDM (Hawkins et al., 2015). Even with this “stacked deck”, the collapsing bounds models appeared favored by BIC for datasets acquired from non-human primates.

In a second experiment, Hawkins et al. (2015) performed the same model fitting analyses but removed trial-variability in drift rate, starting point, and non-decision time for the

[12] See Villarreal et al. (2023) for thought-provoking alternative conceptualizations of complexity that do not appeal to the number of parameters in a model.

collapsing bounds models. These changes much better aligned their specifications of collapsing bounds with the theoretical justification behind this form of the model: recall that Drugowitsch, Moreno-Bote et al. (2012) did not need to invoke any trial-varying parameters to explain the data in their study. Readers are encouraged to compare the "Posterior Model Probability" visualizations in Figures 5 and 6 of Hakwins et al. (2015) to appreciate how drastically these specification changes altered the pattern of results. Primate data that previously strongly supported collapsing bounds now seemed either mixed or to prefer fixed bounds, and human data appeared to favor each form roughly equally (Hawkins et al., 2015, Figure 6). This categorical shift in results underscores the role of alternative models in the model comparison process: they serve as a "control condition", effectively defining the context wherein the performance of a particular model will be assessed. Because most common quantitative metrics return *relative* measures of model performance, it is imperative that alternative models are specified in a way that maximally aligns the comparison with the users' reasoning goals. Further, it is crucial that consumers of model-based research findings keep a critical eye toward the alternative models against which a "primary" model's performance is assessed.

By taking a large, data-driven approach to model comparison, the goal of Hawkins et al. (2015) can be interpreted as identifying the true underlying structure of the decision process. However, as suggested above, we take this aim to be misguided. Each of the competing forms facilitates a different type of reasoning about one's target and reflects different epistemic commitments of the researchers who developed the model. Fixed boundaries with trial-varying parameters provide a statistically robust explanation of a wide range of behavioral data, which broadly reflects reasoning goals in mathematical psychology as a field. Collapsing boundaries without trial-variability provide a normative solution to two-alternative decisions in a heterogeneous environment. They also generate representational resources that can accommodate neurobiological findings not predicted by the DDM: . This form of the model broadly reflects the reasoning goals of theoretical neuroscientists who seek to understand the formal principles linking brain and behavior. Fortunately, it is always possible to compose a model containing a mixture of these components (e.g., Murphy et al., 2016), although integrating trial variability into starting point, drift rate, or non-decision time would sacrifice the normative guarantee of the collapsing boundaries.

We can contrast Hawkins et al.'s (2015) quantitative approach to model comparison to that undertaken by Moran (2015). In a thought-provoking paper, Moran offers a conceptual and mathematical analysis of the optimal decision policy in biased and heterogeneous environments. He begins by showing that Bayes Risk and reward rate are functionally equivalent criteria of optimality, such that any procedure that optimizes one also optimizes the other. This analysis underscores core mathematical properties of the DDM: it is the optimal solution in homogeneous environments and it ceases being optimal in any environment once the trial-varying parameters are incorporated. He further affirms that the optimal solution in heterogeneous environments is collapsing boundaries (Moran, 2015). Then, he conducts an analysis that showcases the utility of formal reasoning: asking how to optimally make decisions in a biased and heterogeneous environment ***if*** the underlying architecture is assumed to be the DDM. In other words, identifying which parameterizations *maximize*—rather than optimize—reward rate with a process known to suboptimal in that environment. Moran (2015) found that the DDM must incorporate environmental bias both as a shift in the starting point and as a bias on the drift rate. This result contradicts those reported by (van Ravenzwaaij et al., 2012), who took a quantitative, simulation-based approach and found that bias only needs to be represented in the starting point parameter. Crucially, it was by performing the formal analysis that Moran (2015) identified an oversight in the simulation results: all of their specifications used a fixed value for the threshold, a parameter that varies both across conditions and individuals and that is crucial for titrating speed-accuracy tradeoffs.

Intermediate summary: different tools for different goals

The past three lessons highlighted how reasoning goals shape every step of model-based research, and underscored the importance of keeping these reasoning goals in mind when evaluating models and model-based findings. In doing so, we were able to identify a principled heuristic for deciding whether to use an extended DDM or a diffusion model with collapsing boundaries: if you wish to reason normatively about decision making in heterogeneous environments, then your model must include collapsing boundaries. The extended DDM was optimized for empirically fitting behavioral data, and thus provides a statistically robust and useful tool for describing how properties of decision making systems change as a function of stimuli or task demands.

Lesson 5: The theory-ladenness of experimental design

In this lesson, we demonstrate one of the potential pitfalls of model-based empirical research: unwitting entrenchment of model representations and/or assumptions into standard experimental design choices. Unlike contemporary concerns with experimental protocols in psychology (e.g., that their lack of theoretical grounding compounds issues of replicability (Guest & Martin, 2021; van Rooij & Baggio, 2021)), this problem emerges when a research area has a relatively long history of theory-guided methodological development. When testing a new theory, researchers have good reason to make experimental design choices that maximally align features of the empirical environment with assumptions and representations at play in their model, as this provides the most direct and sensitive test of predictions (see Simen et al., 2009 and the footnote[13] for an example). These maximally-aligned conditions guide considerations for subsequent research, effectively establishing a paradigm for how to measure behavior such that it can be meaningfully analyzed with the model. While such a paradigm can accelerate theory-driven discovery, it can also obscure the degree to which model assumptions shape researchers' conceptualizations of their targets of study—a problem that becomes compounded over time. We take the integration of expectations into accumulator models as our example of this phenomenon.

Expectations consist of knowledge about the prior probability that one of the two choice options will be correct and/or rewarded on each trial. Understanding how expectations shape choice behavior is a core explanatory target for accumulator models, with theoretical work on the topic pre-dating publication of the original DDM (Edwards, 1965; Link & Heath, 1975). Edwards (1965) showed that, without loss of optimality, expectations can be incorporated into the SPRT as a shift in the starting point reflecting the odds of each of the two options. This decreases the distance between the starting point and decision boundary for the preferred option, leading to faster responses for visual evidence concordant with the expectation. Expectations continued to be modeled in this manner throughout the development of the DDM and other accumulator models (Gold & Shadlen, 2007; Ratcliff & McKoon, 2008), with empirical work largely supporting this model specification (Simen et al., 2009). Importantly, formalizing expectations as a shift in the starting point leads to the

[13] In the context of perceptual decision making, this can mean ensuring that stimuli on each trial are independent and identically distributed (the "i.i.d." assumption) and presenting observers with evidence that changes continuously instead of discretely in time (if positing diffusion instead of sequential sampling).

interpretation that expectations bias the decision process *before* the onset of visual evidence and exert no effects thereafter; that is, they exert a static effect on decisions.

As we saw in previous lessons, optimal parameterizations of the DDM can change considerably when idealizing assumptions are relaxed. The starting point formalization of expectations is no exception. Both Bogacz et al. (2006) and Moran (2015) theoretically demonstrated that when the assumption of environmental homogeneity in the SPRT is relaxed (i.e., when postulating a heterogeneous decision environment), expectations must be modeled *both* as a shift in the starting point *and* an offset on the drift rate in order to maximize reward rate. Despite this being the optimal solution when the DDM is applied in heterogeneous environments, ongoing research often diminishes effects of expectations on the drift rate (e.g., Ratcliff and McKoon, 2008) or posits changes in drift rate and starting point as *competing* explanations (e.g., Mulder et al (2012)) instead of complementary contributions to optimal performance. Importantly, modeling expectations in the drift rate leads to the interpretation that expectations bias the *dynamics* of evidence accumulation, such that evidence will be accumulated faster for decisions with expectation-congruent evidence. However, this is different from expectations exerting *dynamic effects* on evidence accumulation (i.e., those that vary as a function of time within a single decision).

Our argument is that the optimality of the shift-in-starting-point approach has largely precluded considerations of the possibility that expectations exert dynamic effects on decisions. This oversight has propagated primarily through experimental design choices that translate distortive idealizations from the model to the task in order to maximize researchers' ability to detect signatures of optimal decision making behavior. Specifically, the standard practice for manipulating expectations in perceptual decision tasks is to (1) equip observers with perfect knowledge of the prior probability and (2) manipulate that probability on the level of blocks rather than individual trials (Hanks et al., 2011; Kelly et al., 2021; Mulder et al., 2012; O'Connell & Kelly, 2021; Ratcliff & McKoon, 2008; Simen et al., 2009; van Ravenzwaaij et al., 2012). These idealizations are distortive because, in real life, (1) observers *estimate* the prior probability by aggregating across similar previous experiences (Seriès & Seitz, 2013) and (2) the relevant expectation for a particular choice often changes from decision-to-decision (Bornstein et al., 2023). These idealizations can also be considered omissive in the sense that (1) expectation learning is completely omitted from the experimental protocol and (2) decision-level variability in expectations is omitted from the decision environment. Another way to frame this latter point is that the current approach measures behavior in *expectation-homogeneous* decision environments.

As we have seen throughout this case study, relaxing the idealization of environmental heterogeneity greatly changes the form of the optimal model. Specifically, when decision difficulty changes trial-by-trial, the optimal solution is a dynamic adjustment of the decision boundary. In the next lesson, we present a novel theory suggesting an analogous solution for environments where the expectation changes trial-by-trial: dynamic integration of expectations into the decision process (Khoudary et al., 2022).

Lesson 6: the utility of interdisciplinary approaches to theory-guided empirical research

This lesson introduces the accumulator model we built to formalize our theory about how expectations *ought* to be integrated into perceptual decisions *if* those expectations are based on small samples of experience *and* the decision environment is fully heterogeneous (i.e., decision difficulty and expectation vary trial-by-trial). Further, we show how this model synthesizes two disparate empirical findings of time-varying effects of expectations into a common theoretical framework. Although the model has not been shown to maximize reward rate in fully heterogeneous environments, it still offers a normative explanation for dynamic effects: namely, that they are driven by the uncertainty estimation required for statistically optimal inference.

The model assumes that decision makers have learned, through experience, that different stimuli (“cues”) predict one of two possible outcomes (“images”) with unique probabilities. For example, a decision maker could learn that cues A, B, and C predict image X being the correct choice with probabilities 0.8, 0.65, and 0.5. Because there are only two possible outcomes, learning these probabilities for image X also means learning that cues A, B, and C predict image Y with probabilities 0.2, 0.35, and 0.5. The model assumes no loss or distortion of these learned expectations as a function of elapsed time between learning and decision making. One of the core conceptual contributions of the model is to posit that, during cued perceptual decision making, decision makers engage in a process of evidence accumulation in order to set their expectation on each trial. The model further posits that this accumulation process does not terminate once visual evidence becomes available; instead, decision makers continuously sample information from memory in parallel to vision until a decision is made.

As reviewed in Lesson 1, there is precedent within the DDM literature to model memory retrieval as a process of evidence accumulation (Ratcliff, 1978). Further motivation for this mathematical representation comes from the notion of *memory sampling* in accumulator models of value-based decision making (Wang et al., 2021). This family of models posits that when humans and other animals make choices based on the subjective value of two different outcomes (e.g., deciding what to eat for dinner), they sequentially sample past experiences from memory to generate predictions about how valuable each choice is to them (Shadlen & Shohamy, 2016). Our model extends that idea to apply to prior probabilities in perceptual decision making. Crucially, on this formalization, expectations function as an additional source of evidence that is dynamically updated over the course of a decision.

Theoretical and empirical work in multisensory integration has shown that when decisions require combining information from multiple sources of evidence (e.g., vision and hearing), humans perform the statistically optimal procedure of producing a precision-weighted multisensory judgment (Alais & Burr, 2004; Ernst & Banks, 2002; Knill & Saunders, 2003; Landy et al., 2011), where precision is formalized as the variance of the evidence-generating distribution. Our model thus proposes that an internal, dynamically-updating estimate of the precision of each evidence source (memory and vision) drives the integration of expectations with sensory evidence by specifying, at each point in time, how much the decision maker should weigh evidence coming from that source. More directly, we propose that the decision variable in fully heterogeneous perceptual decision environments is a precision-weighted sum of samples from parallel streams of memory and visual evidence. A formal description of the model is provided in the Supplementary Materials.

In addition to providing a normative solution to cued perceptual decision making in fully heterogeneous environments, this model also offers a unifying explanation for previously-reported dynamic effects of expectations. These effects were observed by different research groups using different experimental protocols, and together they offer complementary pieces of evidence in support of a dynamic precision-weighted evidence accumulation process. The first comes from Hanks et al. (2011), who found behavioral and neural evidence that human and non-human primate observers increase their weighting of the prior probability as a function of elapsed time for a given decision. Because this effect occurs later on in the decision process, we will refer to it as a “late effect” of expectations. Importantly, this result was observed in expectation-homogeneous decision environments where observers had perfect knowledge of the prior probability. This context made the finding particularly influential, as it challenged extant theories about the integration of expectations (Moran, 2015; van Ravenzwaaij et al., 2012).

The second result comes from Bornstein et al. (2023), one of the first studies to test how learned expectations influence behavior in expectation-heterogeneous decision environments. First, the authors showed that humans can learn usable expectations through a small amount (~25 trials) of cue-image pairings. Then, they showed that the resultant behavior from cueing these expectations on a trial-by-trial basis was best described by a multi-stage diffusion model, where the terminal state of a memory accumulator set the starting point for vision accumulator on each decision. Finally, Bornstein et al. (2023) found that when the learned cue contained *probabilistic* information about the correct choice (e.g., scene A with 70% probability) and *deterministic* information about the uncertainty in the upcoming visual evidence (e.g., low signal-to-noise ratio), humans sampled from memory in an uncertainty-adaptive manner. A crucial design decision facilitated this observation: introducing inter-stimulus intervals (ISIs) between the onset of the memory cue and visual evidence, the durations of which varied across trials. This allowed Bornstein et al. (2023) to observe that, for as many as 30% of trials where the expectation over accuracy was high and the expectation over signal quality was low, observers opted to make a decision before perceiving any visual evidence at all. Further, this effect was driven by an fMRI measure of memory sampling:

reinstatement of target-specific patterns that were greater over time when the cue signaled low-quality visual evidence. We call this an “early effect” of expectations.

Figure 3 shows that our model can reproduce both of these effects when the relevant differences in the experimental designs are taken into account (i.e., including a delay between cue and sensory evidence for the Bornstein results). Figure 3A shows that a late effect can result from dynamic precision-weighted integration of parallel streams of memory and visual evidence. Each sample of visual evidence contributes to the observer’s estimate of visual evidence precision, which in this case is at chance (50%). As the observer’s certainty about this precision estimate increases over the course of the decision (Figure 3C), it becomes increasingly clear that visual evidence is insufficiently informative to make a choice. The decision variable (in yellow) thus becomes increasingly biased toward the choice predicted by memory (precision=0.8 in this simulation) over the course of sampling, leading to a late effect of expectations. Figure 3B shows that, when provided with probabilistic information about the correct choice and deterministic information about the evidence quality, a dynamic precision-weighted integration process also captures the early effect. During the anticipatory memory sampling period (preceding the dashed vertical line), the observer accumulates evidence from memory which offsets the starting point of the decision variable at visual evidence onset. Then, only a small amount of visual evidence samples are needed to confirm the state of the world predicted by the expectation (choice outcome X with decision quality Y), and the observer can make a decision much earlier on than if they did not have time to sample from memory in anticipation.

Thus, the model that we built to reason about principled solutions to cued decision making in less-idealized environments also offers a unifying explanation for observations that were previously explained by two different theories (i.e., model structures). This lesson thus highlights one of the primary pathways whereby formal modeling advances neuroscience: motivating the development of novel theories by relaxing idealizing assumptions of existing models. As we saw in Lesson 3 (*The value-ladenness of model specification*), there are at least two ways that modelers can engage in the project of building less idealized models. Following Ratcliff et al. (1998, 2001), one can replace mathematical representations required by particular distortive idealizations with representations that are more numerically flexible and enhance the model's ability to fit data from a wider range of decision environments. Alternatively, one can choose which idealizations they wish to relax and develop a mathematical representation that aligns with how they wish to reason about the less-idealized target. In the case of Drugowitsch, Moreno-Bote et al. (2012), the goal was to identify a mathematical structure that optimized reward rate in heterogeneous environments. In our case, the goal was to develop a mathematical representation of a *principled*—not necessarily optimal—procedure whereby dynamic effects of expectations might emerge. We achieved this by first identifying the idealizations that were inadequate for our purpose, and then constraining our mathematical representations to those that have previously found empirical support in neighboring literatures. In doing so, we demonstrate the utility of interdisciplinary approaches to formal theory development in computational cognitive neuroscience.

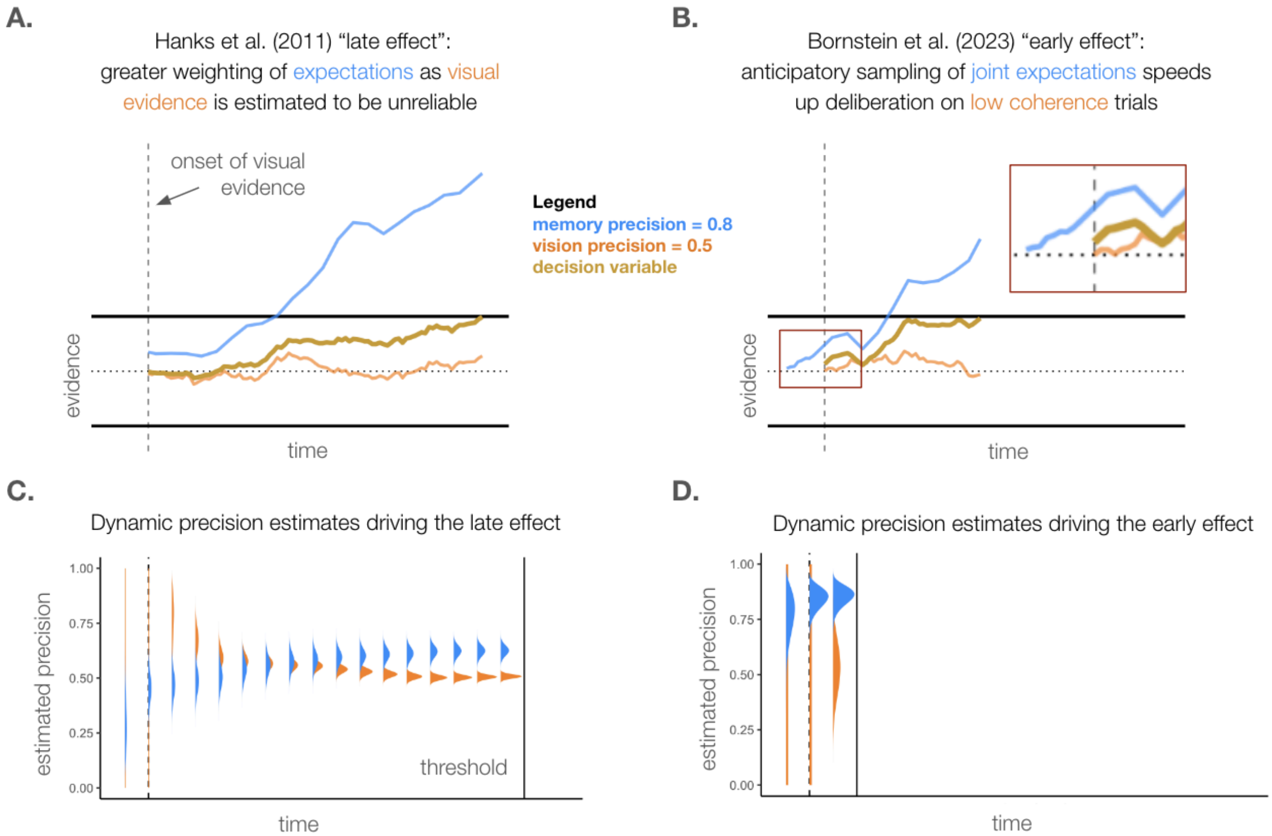

FIGURE 3: Simulated single-trial outputs from the model. In all figures, memory evidence (expectations) is plotted in blue, vision evidence is plotted in orange, and the decision variable is plotted in yellow-brown. (A) Simulated trial where the effect of expectations increases as a function of elapsed decision time, replicating the main result from Hanks et al. (2011). (B) Simulated trial where anticipatory memory sampling expedites deliberation about a highly noisy visual stimulus, replicating a core finding of Bornstein et al. (2023). (C-D) Dynamic evidence precision estimates driving integration of the late and early effects, respectively. Formally, these distributions are the time-evolving posterior Beta distributions quantifying the observer's belief about the variance of each evidence-generating distribution (memory and vision).

Summary and Future Directions

In this article, we summarized some core ideas in philosophy of science, and philosophy of modeling in particular, in the service of developing a philosophical toolkit for computational cognitive neuroscience. The toolkit explicates what is meant by the famous aphorism that “all models are wrong, but some are useful” (Box and Draper, (1987, p. 424), p.424), and shows how the adequacy-for-purpose view of models places a user’s reasoning goals at the heart of the modeling process. We then applied this toolkit to an open question in the accumulator model literature: whether to use a DDM with trial variability in drift rate, starting point, and non-decision time (Ratcliff et al., 2016) or one where these parameters are fixed across but the decision thresholds decrease during the course of a decision (Drugowitsch et al., 2012). We traced the evolution of these forms with an eye toward the reasoning goals behind representational decisions that departed from the original DDM, and demonstrated concretely the utility of a philosophical toolkit for guiding scientific decision-making: if one wishes to reason normatively about perceptual decision making in heterogeneous environments, then one must use the form of the model with collapsing decision boundaries. We then demonstrated how philosophical insights about idealization assisted in identifying an oversight in the literature on cued perceptual decision making, which in turn inspired the development of a novel theory about how expectations dynamically affect perceptual decisions.

There is still much work to be done on the philosophical foundations of computational cognitive neuroscience. We hope that the ideas developed in this article can serve as a starting point for future investigations into the role of reasoning goals in shaping formal models outside the domain of perceptual decision making. We also hope that gaining recognition of how reasoning goals permeate the entire process of model-based empirical research will call readers to reflect on the reasoning goals that motivate their scientific practices, and consider how differences in reasoning goals might be at play when finding oneself in a disagreement about the validity of a particular analytic approach. Altogether, this article aimed to remind readers that computational models—although mathematical in nature—are still scientific tools created by humans whose decisions are inevitably shaped by the values and goals they bring to the problem. Feminist philosophy of science advises that the best way to manage this value-ladenness of representational decision making is to encourage a diverse range of reasoning goals and model types, rather than prescriptively assign higher value to certain goals or specifications relative to others. After all, in the words of Sutton and Barto (2018, p.50), “at present, such representational choices are more an art than a science”. Encouraging (reasonable) creativity in the building of scientific representations is one of the most straightforward ways to accelerate empirical progress in cognitive neuroscience.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology: CB*, *14*(3), 257–262.

Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, *36*(3), 30.

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Attention, Perception & Psychophysics*, *73*(2), 640–657.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 1.

Batchelder, W. H. (2010). Mathematical psychology. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(5), 759–765.

Blohm, G., Kording, K. P., & Schrater, P. R. (2020). A How-to-Model Guide for Neuroscience. *eNeuro*, *7*(1). https://doi.org/10.1523/ENEURO.0352-19.2019

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bornstein, A. M., Aly, M., Feng, S. F., Turk-Browne, N. B., Norman, K. A., & Cohen, J. D. (2023). Associative memory retrieval modulates upcoming perceptual decisions. *Cognitive, Affective & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-023-01092-6

Box, G. E. P., & Draper, N. R. (1987). Empirical model-building and response surfaces. *Wiley Series in Probability and Mathematical Statistics.*, *669*. https://psycnet.apa.org/fulltext/1987-97236-000.pdf

Boyd, N. M., & Bogen, J. (2021). Theory and Observation in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/science-theory-observation/

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience*, *13*(1), 87–100.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *12*(12), 4745–4765.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*(6128), 95–98.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological Review*, *100*(3), 432–459.

Cartwright, N. (1983). *How the Laws of Physics Lie*. OUP Oxford.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*(6), 693–702.

Cohen, J. D., Daw, N., Engelhardt, B., Hasson, U., Li, K., Niv, Y., Norman, K. A., Pillow, J., Ramadge, P. J., Turk-Browne, N. B., & Willke, T. L. (2017). Computational approaches to fMRI analysis. *Nature Neuroscience*, *20*(3), 304–313.

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems*

Research, *27*, 42–49.

Dayan, P., & Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: different models but the same conclusions about psychological processes? *Psychonomic Bulletin & Review*, *18*(1), 61–69.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(11), 3612–3628.

Eberhardt, F., & Danks, D. (2011). Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models. *Minds and Machines*, *21*(3), 389–410.

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology*, *2*(2), 312–329.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666.

Forstmann, B. U., & Turner, B. (Eds.). (2024). *An Introduction to Model-Based Cognitive Neuroscience*. Springer International Publishing.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-Based Cognitive Neuroscience: A Conceptual Introduction. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An Introduction to Model-Based Cognitive Neuroscience* (pp. 139–156). Springer New York.

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279.

Frazier, P., & Yu, A. J. (2007). Sequential hypothesis testing under stochastic deadlines. *Advances in Neural Information Processing Systems*, 465–472.

Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/models-science/

Gluth, S., & Meiran, N. (2019). Leave-One-Trial-Out, LOTO, a general approach to link single-trial parameters of cognitive models to neural data. *eLife*, *8*. https://doi.org/10.7554/eLife.42607

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16.

Gold, J. I., & Shadlen, M. N. (2002). Banburismus and the brain: decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, *36*(2), 299–308.

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

Grahek, I., Schaller, M., & Tackett, J. L. (2021). Anatomy of a Psychological Theory: Integrating Construct-Validation and Computational-Modeling Methods to Advance Theorizing. *Perspectives on Psychological Science: A Journal of the Association for Psychological*

Science, *16*(4), 803–815.

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(4), 789–802.

Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., & Brody, C. D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*, *520*(7546), 220–223.

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(17), 6339–6352.

Harvard, S., & Winsberg, E. (2022). The Epistemic Risk in Representation. *Kennedy Institute of Ethics Journal*, *32*(1), 1–31.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(6), 2476–2484.

Howard, M. W., & Kahana, M. J. (2002). A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Jones, M. R. (2005). IDEALIZATION AND ABSTRACTION: A FRAMEWORK. In *Idealization XII: Correcting the Model* (pp. 173–217). Brill.

Kelly, S. P., Corbett, E. A., & O’Connell, R. G. (2021). Neurocomputational mechanisms of prior-informed perceptual decision-making in humans. *Nature Human Behaviour*, *5*(4), 467–481.

Khodadadi, A., & Townsend, J. T. (2015). On mimicry among sequential sampling models. *Journal of Mathematical Psychology*, *68-69*, 37–48.

Khoudary, A., Peters, M. A. K., & Bornstein, A. M. (2022). Precision-weighted evidence integration predicts time-varying influence of memory on perceptual decisions. *Cognitive Computational Neuroscience*. https://aaron.bornstein.org/cv/pubs/2022_kpb_ccn.pdf

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, *43*(24), 2539–2558.

Kording, K., Blohm, G., Schrater, P., & Kay, K. (2018). *Appreciating diversity of goals in computational neuroscience*. https://doi.org/10.31219/osf.io/3vy69

Korzybski, A. (1933). *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. International Non-Aristotelian Library Publishing Company.

Krueger, P. M., van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P., & Cohen, J. D. (2017). Evidence accumulation detected in BOLD signal using slow perceptual decision making. *Journal of Neuroscience Methods*, *281*, 21–32.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Wiley.

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. In J. Trommershauser, K. Kording, & M. S. Landy (Eds.), *Sensory Cue Integration*. Oxford University Press.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust Modeling in Cognitive Science. *Computational Brain & Behavior*, *2*(3), 141–153.

Levenstein, D., Alvarez, V. A., Amarasingham, A., Azab, H., Chen, Z. S., Gerkin, R. C., Hasenstaub, A., Iyer, R., Jolivet, R. B., Marzen, S., Monaco, J. D., Prinz, A. A., Quraishi, S., Santamaria, F., Shivkumar, S., Singh, M. F., Traub, R., Nadim, F., Rotstein, H. G., & Redish, A. D. (2023). On the Role of Theory and Modeling in Neuroscience. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *43*(7), 1074–1088.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*(1), 77–105.

Longino, H. (1998). *Values and objectivity*. na.

Love, B. C. (2020). Model-based fMRI analysis of memory. *Current Opinion in Behavioral Sciences*, *32*, 88–93.

Moran, R. (2015). Optimal decision making in heterogeneous and biased environments. *Psychonomic Bulletin & Review*, *22*(1), 38–53.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(7), 2335–2343.

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications*, *7*, 13526.

Navarro, D. J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Computational Brain & Behavior*, *2*(1), 28–34.

O'Connell, R. G., & Kelly, S. P. (2021). Neurophysiology of Human Perceptual Decision-Making. *Annual Review of Neuroscience*, *44*, 495–516.

O'Connell, R. G., Shadlen, M. N., Wong-Lin, K., & Kelly, S. P. (2018). Bridging Neural and Computational Viewpoints on Perceptual Decision-Making. *Trends in Neurosciences*, *41*(11), 838–852.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53.

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*(Pt B), 59–64.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376–404.

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science*, *87*(3), 457–477.

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in perceptual and value-based decision making. *Neuron*, *82*(3), 709–720.

Portides, D. (2021). Idealization and abstraction in scientific modeling. *Synthese*, *198*(24), 5873–5895.

Press, C., Yon, D., & Heyes, C. (2022). Building better theories. *Current Biology: CB*, *32*(1), R13–R17.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *The Behavioral and Brain Sciences*, *41*, e223.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

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., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356.

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.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2021). Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(4), 725–743.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *22*(21), 9475–9489.

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-020-00938-0

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, *6*(2), 461–464.

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in Human Neuroscience*, *7*, 668.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936.

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(6), 1865–1897.

Smaldino, P. E. (2020). How to Build a Strong Theoretical Foundation. *Psychological Inquiry*, *31*(4), 297–301.

Stanford, P. K. (2006). *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford University Press.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to Analysis in Model-based Cognitive Neuroscience. *Journal of Mathematical Psychology*, *76*(B), 65–79.

van Ravenzwaaij, D., Mulder, M. J., Tuerlinckx, F., & Wagenmakers, E.-J. (2012). Do the dynamics of prior information depend on task context? An analysis of optimal performance and an empirical test. *Frontiers in Psychology*, *3*, 132.

van Rooij, I., & Baggio, G. (2021). Theory Before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(4), 682–697.

van Rooij, I., & Blokpoel, M. (2020). Formalizing Verbal Theories. *Social Psychology*, *51*(5), 285–298.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*(1), 37–58.

Villarreal, M., Etz, A., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological Review*, *130*(4), 853–872.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, *16*(2), 117–186.

Wald, A., & Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *Annals of Mathematical Statistics*, *19*(3), 326–339.

Wang, S., Feng, S. F., & Bornstein, A. M. (2021). Mixing memory and desire: How memory reactivation supports deliberative decision-making. *Wiley Interdisciplinary Reviews. Cognitive Science*, *13*(2), e1581.

Weber, J., Solbakk, A.-K., Blenkmann, A. O., Llorens, A., Funderud, I., Leske, S., Larsson, P. G., Ivanovic, J., Knight, R. T., Endestad, T., & Helfrich, R. F. (2024). Ramping dynamics and theta oscillations reflect dissociable signatures during rule-guided human behavior. *Nature Communications*, *15*(1), 637.

Wimsatt, W. C. (1987). False Models as Means to Truer Theories. In M. H. Nitecki & A. Hoffman (Eds.), *Neutral Models in Biology* (pp. 23–55). Oxford University Press.

Winsberg, E., & Harvard, S. (2024). Scientific Models and Decision Making. In *Elements in the Philosophy of Science*. Cambridge University Press.

Wu, J. (2023). Better than Best: Epistemic Landscapes and Diversity of Practice in Science. *Philosophy of Science*, 1–11.