**Title:** Reasoning goals and representational decisions in computational cognitive neuroscience: lessons from the drift diffusion model

**Authors:** Ari Khoudary[1-3], Megan A. K. Peters[1-5*], 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
4. Department of Logic and Philosophy of Science, University of California, Irvine
5. Brain, Mind, and Consciousness Program, Canadian Institute for Advanced Research

*These authors contributed equally.

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

**List of abbreviations:**

- DDM: drift diffusion or diffusion decision model
- RT: response or reaction time
- EEG: electroencephalography
- fMRI: functional magnetic resonance imaging
- MDP: Markov decision process
- SPRT: sequential probability ratio test
- RRO: reward rate optimality
- BIC: Bayesian information criterion
- AIC: Akaike information criterion

**Citation diversity statement**

In this paper, we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. To assess the degree to which we achieved these goals with respect to gender, race, and ethnicity, we utilized an open-source software package (https://github.com/dalejn/cleanBib; Zhou et al., 2022) that evaluates our reference list and probabilistically assigns gender and racial/ethnic identities to the first and last authors of cited work. This method is limited in that (a) it cannot account for intersex, non-binary, or transgender people, (b) names, pronouns, and social media profiles used to construct the databases do not always accurately capture gender identity, (c) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the racialization or ethnicization of their names, and (d) names and Florida Voter Data used to make the predictions may not be indicative of racial/ethnic identity. Keeping these limitations in mind, we report results of this analysis in order to raise and maintain awareness about social imbalances in scientific research.

First, the software package obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2022). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 7.95% woman(first)/woman(last), 13.5% man/woman, 9.98% woman/man, and 68.57% man/man. Next, the software package obtained predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color (Ambekar et al., 2009; Chintalapati et al., 2018). By this measure (and excluding self-citations), our references contain 5.1% author of color (first)/author of color(last), 13.54% white author/author of color, 14.28% author of color/white author, and 67.08% white author/white author.

**Acknowledgements:** We thank Ainsley May for several discussions and comments on an early draft, along with Kevin O'Neil, Dale Zhou, Felipe De Brigard, and three anonymous reviewers for helpful feedback on later drafts. 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. A portion of this project was presented at the Deep South Neuroscience and Philosophy workgroup where the authors engaged in a number of helpful conversations. This research was supported by NIMH T32 MH119049 to A.K. and partially supported by both the Canadian Institute for Advanced Research (Fellowship in Brain, Mind, & Consciousness to M.A.K.P.) and the Air Force Office of Scientific Research (award number FA9550-20-1-0106 to M.A.K.P.).

**Abstract**

Computational cognitive models are powerful tools for enhancing the quantitative and theoretical rigor of cognitive neuroscience. It is thus imperative that model users—researchers who develop models, use existing models, or integrate model-based findings into their own research—understand how these tools work and what factors need to be considered when engaging with them. To this end, we have developed a philosophical toolkit that address core questions about computational cognitive models in the brain and behavioral sciences. Drawing on recent advances in philosophy of modeling, we highlight the central role of model users' *reasoning goals* in the application and interpretation of formal models. We demonstrate the utility of this perspective by first offering a philosophical introduction to the highly popular drift diffusion model (DDM) and then providing a novel conceptual analysis of a long-standing debate within that model's literature. Contrary to most existing work, our analysis suggests that the two models implicated in the debate offer complementary—rather than competing—explanations of speeded choice behavior. We achieve this by first explicating the role of optimality in model-based explanations, and then demonstrating how the two different models reflect different commitments to optimality explanations. We use these insights to offer a principled heuristic for when to use one form versus the other, before concluding with a critical appraisal of reasoning goals in formal model comparison. Altogether, we demonstrate the conceptual and practical utility of philosophy for inspiring new directions in brain and behavioral research.

*Keywords*: model specification, optimality, idealization, explanation

**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

## 1. Introduction

Formal theorizing via computational modeling is often lauded lauded as a solution to meta-scientific challenges in the brain and behavioral sciences (Forstmann et al., 2011; Grahek et al., 2021; Guest & Martin, 2021; Muthukrishna & Henrich, 2019; Press et al., 2022; Robinaugh et al., 2021; Turner et al., 2019). While formal computational models indeed have much to offer toward this end, we also recognize them as incredibly powerful and flexible tools that—when misused—can create even worse problems than they were initially applied to solve. Further, subfields of psychology that rest upon decades of formal modeling work face their own meta-scientific challenges that—by their nature—cannot be resolved by use of a model alone. In this article, we aim to demonstrate the utility of philosophical research for mitigating and/or resolving such meta-scientific challenges. To do this, we have compiled insights from foundational and recent work in philosophy of modeling to develop a "philosophical toolkit" for computational cognitive modeling. We take as an applied example sequential sampling models of choice-reaction time, with a focus on the drift diffusion or diffusion decision model (DDM).

A number of factors motivated our goal to focus on sequential sampling models. First, this family of models, and the DDM in particular, are highly prominent in computational cognitive (neuro)science. This fact, together with the aforementioned goal of integrating formal models into psychological research, has created demand for a number of different software packages that increase the accessibility of these formal modeling tools (e.g., the "EZ-DDM" by Wagenmakers et al., 2007 and the "PyDDM" by Shinn et al., 2020). A philosophical and conceptual analysis of how these models work and what factors need to be considered when using them both supports newer researchers' entry to this research area, and helps mitigate against confused applications or erroneous conclusions that might arise in the course of learning how to use these tools.[1] We further hope that our toolkit offers experienced researchers an opportunity to reflect on their own modeling practice and the goals that shape it. Finally, we aim to contribute to the philosophical literature a rich case study about a family of models that have been almost entirely overlooked by the discipline. Whereas the merits and limitations of, e.g., connectionist and Bayesian models have received substantive philosophical attention, the DDM has has only featured marginally in recent work, primarily as an example of cognitive models more broadly (Figdor, 2018; Drayson, 2020; Gamboa, 2024). Thus, we hope

[1] This article is not intended as a guide for software usage. Conceptual and practical considerations related to implementing computational cognitive models can be found in Cooper & Guest (2014), Lee & Wagenmakers (2014), Lin & Strickland (2020), and Wilson & Collins (2019)

this introduction to the DDM and its broader model family can serve as a starting point for future philosophical research about the role(s) of these models in computational cognitive (neuro)science.

We first provide readers with a collection of philosophical tools for thinking about the practice and products of computational cognitive neuroscience research. We then demonstrate the utility of such a "toolkit" in two ways: offering a philosophical introduction to the DDM, and then offering a novel conceptual analysis of a long-standing debate about the form of decision boundaries in the DDM. This analysis is also one of the first contributions to the debate that does not aim to argue in favor of one form versus the other. Rather, we use an analysis of the reasoning goals motivating each "side" in the debate to (1) argue that the forms are complementary rather than competing, and (2) offer readers a principled heuristic for deciding when to use either form. In doing so, we critically review the model selection process and characterize the role that optimality considerations play in driving model-based scientific research.

## 2. Philosophical tools for thinking about computational cognitive models

The term "model" is used to refer to a number of related but functionally distinct objects involved in scientific research. One thing common to most models is that they are *representations* of a **target**—some phenomenon in the natural world—that make it easier for a human (or group of humans) to reason about that target (e.g., van Rooij, 2022). This article focuses on models that are abstract representations of the latent entities and processes generating an organism's behavior, often called *cognitive* or *process* models. Scientists use these models both to predict how an organism (or agent) will behave in response to experimental manipulations, and to explain why a particular manipulation induced the behavioral response that it did. In some cases, these models are *also* tasked with predicting and explaining changes in neural activity related to changes in behavior. Finally, scientists sometimes use these models simply as measurement devices for quantifying individual differences in latent properties or processes that are relevant for a broader type of explanation. Thus, the *same* model can be used in the service of different goals (prediction, explanation, measurement) directed toward targets at different conceptual/spatiotemporal scales (cognitive processes and/or neural activity). This multiplicity is both what makes models so useful for scientific practice ***and*** is a primary factor contributing to confusion and debates about how best to use models in scientific research. Reviewing some philosophical research concerned with the questions of what models are and how they work can help mitigate these confusions while also offering a salve against the "existemic"[2] concerns that model-based cognitive (neuro)science research can so frequently incur.

[2] We coin "existemic" to refer to feelings of existential dread brought about by reflecting on the epistemic limitations of particular ways of knowing about the world.

Based on the uses described above, the types of models we consider in this article are simultaneously (i) abstract representations of latent entities and processes that generate an organism's behavior, (ii) theories that predict and explain changes in the organism's behavior, and (iii) devices for measuring the latent entities and processes represented in the model. Based on the topics discussed in our case study, we focus our attention on properties (i) and (ii), with a particular focus on how these properties relate to model-based explanations. Following Weisberg (2013), we consider these models *computational* if their theoretical content appeals to transition rules or an algorithm (i.e., a series of steps specifying how an initial state is transformed into an output).[3] Importantly, these properties do not apply to all of the models used in the behavioral and brain sciences, and certainly not to all of the objects reasonably considered scientific models.[4] But they do aptly characterize the use of sequential sampling models in contemporary computational cognitive neuroscience (see also Forstmann et al., 2016; and Turner et al., 2019 for additional examples). The rest of this section introduces the philosophical research that has clarified our own thinking about these models. Readers primarily interested in learning about the DDM can skip ahead to section 3.

*2A: Tools for thinking about models as representations*

Our toolkit begins with a survey of some philosophical literature concerned with the representational nature of formal models, with a focus on concepts most relevant to our case study in Section 4. A large body of research spanning philosophy of science, philosophy of mind, and philosophy of language is devoted to explicating the concept of representation; Frigg & Nguyen (2017) offer a helpful overview for readers interested in the role of representations in science. For our purposes, it suffices to say that a model becomes a representation of a target when a **model user**—a human who builds or interprets the model—***decides*** that they are going to map components and processes of the target onto components and processes of the model, ultimately with the goal of using properties of the model to reason about corresponding properties of the target (Winsberg & Harvard, 2024). On this view, a model represents a target by "standing in" for the target in a way that permits the model user to reason about their target on the basis of interacting with the model. This process has been called *surrogative reasoning* in the philosophical literature. When coining this term, Swoyer (1991) offers a helpful example: "By using numbers to represent the lengths of physical objects, we can represent facts about

[3] Weisberg (2013) considers computational models a special case of mathematical models because of how scientists use them to explain. On his account, mathematical models use sequences of states or an equilibrium as the explanation, whereas computational models use the *procedure* that specifies how an input is transformed into an output.

[4] For example, some models are *concrete* representations, like scale models used in architectural engineering and rodent models used in translational neuroscience research. Other models are abstract representations that compactly summarize many observations about a target process, or describe how a target responds when it is intervened upon, rather than comprising theories that can be used to explain why the target responded how it did. These types of models are often called "descriptive", "phenomenological", or "data" models.

the objects numerically, perform calculations of various sorts, then translate the results back into a conclusion about the original objects. In such cases, we use one sort of thing as a surrogate in our thinking about another" (p. 87).

Computational cognitive models use mathematical objects (numbers, distributions, vectors, etc.) and equations to represent the cognitive entities and operations that generate an organism's behavior. Scientists thus use these mathematical representations as *surrogates* for thinking about what processes "under the hood" are driving behavior. Recent work in philosophy helpfully distinguishes between a model's ***structure*** and its ***construal***: the first being the mathematical objects and equations comprising the representation, and the second being how those abstract objects are meant to be mapped onto objects and processes in the physical world, respectively (Weisberg, 2013; Andrews, 2021). The simple example above demonstrates how construals are necessary for interpreting the structure of a model. If a user was presented only with the set of numbers that correspond to the length of physical objects, but was not told what properties of the physical world those numbers are correspond to, they would (1) have little to no guidelines for constraining the types of mathematical reasoning appropriate to perform with that representation and (2) have no way of translating the results of even simple mathematical operations on the representation into actions in the real world. Construals thus function as a "bridge" between the mathematical domain where we construct and manipulate a representation of the target and the physical domain wherein we intervene upon and collect measurements from it. As such, they are central to the practice of model-based scientific research.

On Weisberg's (2013) account, a model's construal consists of three components: assignment, scope, and fidelity criteria. The *assignment* of a model explicitly specifies how parts of the target system are mapped onto parts of the model; in other words, it states what each variable in the model is supposed to correspond to inside an organism's head. The *scope* of a model specifies which aspects of the target the model aims to represent (and, by extension, which aspects it does *not* aim to represent). Considering a model's intended scope is essential for determining which types of measurements from the target are meaningful for assessing the performance of the model. The final components of a construal, *fidelity criteria*, are the benchmarks that model users reference in order to determine whether they have a "good" model of their target. Our case study in Section 4 demonstrates how differences in fidelity criteria between different subgroups of researchers using the DDM have resulted in current "competing" forms of the model.

A point we wish to emphasize is that both a model's structure and its construal are the products of *decisions* made by users who build and apply the model to data. And just like in cognitive decision-making, these ***representational decisions***—determining *what* properties of the target to include in the model and *how* to represent those properties mathematically (Harvard and Winsberg, 2022)—are subject to variation across users in different subfields, across users within the same subfield, and even within a single user applying the same model in different contexts. We argue that this variability reflects the multifaceted roles that models

play in scientific reasoning, and that understanding the factors contributing to variability in representational decision-making is essential for informed, critical engagement with model-based scientific research. Echoing recent work in philosophy (Danks, 2015; Potochnik & Sanches de Oliveira, 2020; Weisberg, 2013; Winsberg & Harvard, 2024) and computational neuroscience (Blohm et al., 2020; Kording et al., 2018), we aim to demonstrate how model users' ***reasoning goals***—which aspect(s) of the target they aim to reason about and how they wish to perform that reasoning via the model—are primary drivers of variability in representational decisions. Because representational decisions determine both a model's structure *and* its construal (which itself specifies fidelity criteria), this position implies that reasoning goals shape the model-based research process all the way down to quantitative model comparison procedures; we provide concrete examples in Section 4.

One might worry that permitting even quantitative model comparison procedures to vary according to a user's goals might be too flexible of a philosophy of modeling for the practicing scientist. On our view, however, this position is an inevitable consequence of the fact that most—if not all—formal models are *incomplete* representations of their target systems. Because this property of models makes them technically false with respect to their target (Frigg & Hartmann, 2020; Wimsatt, 1987), model users cannot rely simply on the "truth" or falsity of models in order to select the best among them. A prominent line of reasoning in philosophy thus suggests that models be evaluated by their *adequacy for purpose* rather than *verisimilitude* (i.e., true or accurate representation of the target) alone (e.g., Parker, 2020).[5] By positing that the goal of modeling is to identify representations that are *useful*, rather than strictly-speaking *true*, the adequacy-for-purpose view further emphasizes the central role of reasoning goals in model-based science: they define the purpose against which a model's adequacy is evaluated. Because different models are built and/or applied in the service of different reasoning goals, it is desirable to allow fidelity criteria to vary according to the reasoning goal being pursued. The task of model-based science thus becomes identifying which representations are adequate for which purposes, and, at a higher-level, identifying the reasoning goals that are most useful for making progress on particular scientific questions.

Further motivation for the adequacy-for-purpose view comes from the heavy use of idealization in formal models of complex systems (Cartwright, 1983; Potochnik, 2018). When scientists build idealized representations of their target systems, they build representations that intentionally *mis*represent features of their target in order to make reasoning about it easier. It can be useful to distinguish between *omissive* and *disortive* idealizations: those that remove certain properties of the target from the representation (e.g., choosing not to represent neural dynamics in a cognitive model) and those that represent known properties in a way that is known to be inaccurate (e.g., assuming that observers have perfect knowledge of the environment), respectively. These complementary forms of idealization—at play in nearly all

[5] Readers might already be familiar with this position on the basis of the popular aphorism "all models are wrong, but some are useful" (Box & Draper, 1987, p. 424).

formal models—permit users to build representations that "selectively attend" to components of the objects that the model user wishes to reason about (Portides, 2021). In doing so, the model reduces the complexity of the target such that reasoning about it becomes more tractable for the model user. The task of modeling, again, becomes identifying the degree and type(s) of idealizations that are most useful for one's purposes. The rest of this article aims to provide tools for thinking about this set of decisions.

*2B: Tools for thinking about models as explanations*

Equipped with some tools for thinking about models as representations, we next turn to research in computational neuroscience and philosophy to consider the types of reasoning goals enabled by formal models. Kording et al. (2018) helpfully identify twelve different reasoning goals most commonly pursued in computational neuroscience, while noting that "it is impossible to produce an exhaustive list" (p.3). Wimsatt (1987) also proposes "Twelve things to do with false models" (pp. 7-8) based on his work with formal models in engineering. That these two lists minimally overlap both reflects how broad the space of possible reasoning goals is in model-based science, and highlights the utility of integrating insights developed independently in neuroscience and philosophy. Based on the argument we develop in the case study, we will focus our discussion here on reasoning goals related to *explanation.*[6] First we discuss *how* formal models give explanations of empirical targets and then contrast different *types* of explanations formal models can give.

Foundational work in the philosophy of scientific explanation distinguishes between an observation or phenomenon that scientists aim to explain (i.e., an *explanandum*) and the explanation that scientists give of it (i.e., the *explanans*; Hempel & Oppenheim, 1948).[7] When a formal model is used as an explanation for some explanandum, the model's *structure* functions as the explanans (Weisberg, 2013). In other words, the target's behavior is explained by virtue of the structure of its formal/mathematical representation. Weisberg (2013) argues that, in computational models, this structure is comprised of the procedures (or algorithms) that specify how an input state is transformed into an output, a position we adopt here as well. On this account, computational cognitive models explain the behavior of their targets by relating (or "mapping") changes in observed behavior onto changes in the parameter values and/or configurations of the latent procedures represented in the model's structure. This mapping can

[6] Explanation is one of the oldest topics in philosophy of science and thus cannot be exhaustively covered here. Helpful introductions to ongoing philosophical work on explanation in the cognitive and brain sciences can be found in Kaplan (2017) and a special issue edited Colombo & Knauff (2020).

[7] Whether an explanandum is different from a model's target depends on what the user construes as the target of a particular model. In the case when the target is a particular pattern of observations observed in particular experimental contexts, there is no meaningful difference between a target and an explanandum. However, when the target of a model is a general (neuro)cognitive process (e.g., speeded two-alternative decision making), then particular observations of the target in particular contexts constitute different explananda that the model is tasked with unifying into a single generative formal structure.

be achieved both by simulating the target's behavior on different parameter values or configurations of the model's structure (i.e., "simulation") and/or by identifying parameter values of components of the structure that maximize the likelihood of observing a particular set of observations from the target (i.e., "model fitting").

At the heart of both these approaches to model-based explanation is the notion of a model "capturing" properties of its target. Weisberg (2013) proposes that scientists use two different types of *fidelity criteria* to assess whether—and in what ways—a model "captures" features of its target. *Dynamical* fidelity refers to the quantitative similarity between measurements taken of the target and numerical estimates generated by the model (e.g., its numerical "goodness-of-fit"), whereas *representational* fidelity refers to how closely a model's structure matches the causal structure of the real-world phenomenon (Weisberg, 2013). For our purposes, the relevant sense of "causal structure" is synonymous with the "type of explanation" a user wishes to attain by reasoning with the model, a topic we treat in the next paragraph. We will call our notion "explanatory fidelity" to differentiate it from Weisberg's (2013) causal notion of representational fidelity. Importantly, these notions of fidelity are defined independently of the methods used to evaluate them, meaning that model users have to make representational decisions in order to evaluate the fidelity of their model: deciding *what* kind of criteria are most appropriate for their overall goals and *how* they wish to evaluate their model with respect to those selected criteria.

Our case study in Section 4 demonstrates the utility of fidelity criteria as tools for thinking about model-based research. In particular, we highlight how the two "competing" forms of the DDM (further described in Section 3) reflect differences in fidelity criteria between communities of scientists using the model, and argue that these differences reflect the different explanatory aims of each community. To do this, we make use of a popular taxonomy that distinguishes *what*, *how*, and *why* approaches to giving explanations (Dayan & Abbott, 2005; Ross & Woodward, 2023). The "what" approach focuses on characterizing how the target behaves under various circumstances, and is commonly thought to be the goal of *descriptive* models that are not generally thought to be explanatory. The "how" approach focuses on decomposing a target's behavior into its constituent parts and processes, and can be said to explain the behavior of the target by virtue of those constituent components; this is commonly considered as the goal of *mechanistic* models in neuroscience (Craver, 2007; Dayan & Abbott, 2005). The "why" approach focuses on identifying properties of the target that require it to behave in particular ways under particular circumstances; this is how we understand the goal of *normative* models in neuroscience (Anderson, 1990; Dayan & Abbott, 2005).

In computational neuroscience, labels from the above taxonomy are commonly used to refer to the entirety of a model's structure. Our case study, however, demonstrates that these properties can also be applied to *individual components* of a model's structure, such that a particular stage in the input-transformation process can be normative (or mechanistic, or descriptive) even if the structure as a whole is not. Although this perspective might complicate usage of the taxonomy, it reflects growing agreement among philosophers that not all

components of a model need to contribute to the user's ultimate purpose in the same way (Weisberg, 2013). Some model components, for example, are included simply because the fitting process would be intractable without them; they are thus included for practical reasons. Other components might be included because a user thinks that component is important for their ability to reason about their target with the model, but particular details about the component might be irrelevant for how it ultimately figures into the primary reasoning goal (e.g., the non-decision term in the DDM). Ideally, model users are explicit about how they intend for each component of their model to be construed with respect to their target, and models are constructed in such a way that these "convenience variables" are not central to the explanation a particular model provides. But because this is not always the case, it is crucial that users keep the heterogeneity of justifications for representational decisions in mind when engaging with the products of model-based research.

*2C: Tools for thinking about models as normative explanations*

This final section of the toolkit provides a brief conceptual analysis of the practice of normative modeling. Our motivations for doing so are (1) to equip readers with the necessary resources for engaging with key topics in the case study, (2) to articulate our own perspective on the role of normative models in computational cognitive neuroscience, and (3) to give readers tools for thinking about a concept that figures heavily in both scientific and meta-scientific debates about perceptual decision making (e.g., Rahnev and Denison, 2018).

Like the term "model", the term "normative" has a number of different but closely related meanings. In general, a normative statement is one that prescribes a particular course of action in a particular scenario. These statements (or logical attitudes) thus take the form: ***if*** your goal is *X*, then you ***ought*** to *Y*. Norms thus create a standard or benchmark for evaluating behaviors as good/correct or bad/incorrect. Crucially, as the logical form makes clear, the accuracy or "goodness" of a particular action is determined *relative* to a particular goal. Normative models in computational cognitive science leverage this property to offer explanations about *why* the target behaved as it did. Users achieve this goal by formally specifying

(1) an *objective function*: the problem the target is trying to solve (e.g., choosing between two options in the shortest possible amount of time)

(2) the *environment* in which it is tasked with solving that problem (e.g., one with or without feedback after each choice)

(3) the *procedure* that a target uses to solve that problem (e.g., sequential sampling to a fixed threshold).

Optionally, users can add a fourth component capturing any *constraints* they wish to impose on the procedure (e.g., leakage or loss of accumulated evidence over time).[8] This

[8] Identifying normative solutions to formal problems that represent different types of constraints a user wishes to incorporate into their model is the premise of the *resource-rational* approach to cognitive modeling (Lewis et al., 2014; Lieder & Griffiths, 2019).

comprehensive formal representation allows users to identify the logical or mathematical limit of the specified target's ability to solve the specified problem in the specified environment, i.e., *optimal* behavior on the experimental task. A common explanatory approach is thus to state that the target exhibited a particular pattern of behavior because it is the optimal solution to the formally-specified problem.

This type of optimality explanation is made possible by *formal frameworks*, which can be defined a set of axioms/postulates (i.e., statements accepted as true), formal/mathematical objects, and rules constraining the relationships among those objects. In other words, formal frameworks offer model users a "grammar" for expressing questions and answers in a format that permits quantitative assessment (Guest & Martin, 2021; Press et al., 2022). Sutton & Barto's (2018) textbook on reinforcement learning offers a helpful example[9]:

> The MDP [Markov decision process] framework is a considerable abstraction [idealization] of the problem of goal-directed learning from interaction. It proposes that whatever the details of the sensory, memory, and control apparatus, and whatever objective one is trying to achieve, any problem of learning goal-directed behavior can be reduced to three signals passing back and forth between an agent and its environment: one signal to represent the choices made by the agent (the actions), one signal to represent the basis on which the choices are made (the states), and one signal to define the agent's goal (the rewards). (p. 50, bracketed text added)

The above quotation also highlights the central role that idealization plays in normative modeling. Because axioms of formal frameworks primarily function to promote mathematical expressivity, modelers often have to make a number of distortive assumptions about their targets in order to build normative models of its behavior. A common example is the widely-used assumption that agents have perfect knowledge of the statistical properties of their environment (e.g., Wald & Wolfowitz, 1948). This assumption exemplifies a distortive idealization of the target because modelers do not often think that the target actually has this perfect knowledge—either in the real world or in the experiment—but still represent their target in this way because it allows them to compute a normative benchmark for performance on the task. Our case study demonstrates two approaches users of the DDM have taken when the assumptions of existing models are inadequate for their reasoning goals.

The pervasiveness of idealization in normative models complicates the question of how to interpret the *failure* of a normative model to explain its target. Discussions concerning the

[9] Some common formal frameworks in psychology and neuroscience include signal detection theory, sequential analysis, information theory, Bayesian inference, and Markov decision processes.

"Great Rationality Debate" in economic decision making[10] suggest three classes of possible interpretations: (1) that the target is generally suboptimal on this task, (2) that more empirical data are needed to understand what factors drive suboptimal performance on this task, or (3) that the formal definition of optimality is inadequate for explanatory purposes. A complementary debate has recently begun in the study of perceptual decision making, with researchers questioning broadly-scoped claims about the *optimality* of perceptual decisions. Rahnev and Denison (2018) comprehensively review evidence demonstrating that humans behave suboptimally on every type of task where their behavior has been considered optimal. The authors do not suggest that the data license an inference that humans are perceptually suboptimal (option 1 from the rationality debate), but instead that the field drop its emphasis on optimality in favor of building detailed models that capture *all* aspects of the perceptual decision process (a model-focused version of option 2 from the rationality debate).

Rahnev and Denison (2018) identify two conceptual challenges of normative models that motivate their suggestion. The first pertains to the variability of optimality definitions across model specifications, and the second pertains to the utility of optimality claims for predicting and explaining behavior. At a broader level, Rahnev and Denison's (2018) critique can be understood as a much-needed reminder for the field that optimality is not a property of targets that can be "discovered" using scientific methods because it is not something that exists independently of the mathematical models that are used to define it. This is exemplified by optimality claims being inherently *contextual* in nature (i.e., dependent on formal specifications of the objective function and the environment) and based on *idealized* representations of the target. Our suggestion, in line with option 3 from the rationality debate, is that users leverage these properties to build normative models that better align with their goals in reasoning about a target. Exploring normative solutions to formally-specified questions not yet addressed in the literature is a principled and straightforward way to advance theorizing in neuroscience, and often motivates the creation of novel experiments that measure the target's behavior in these differently-idealized settings (e.g., Harhen and Bornstein, 2023; Khoudary et al., 2022). Our case study demonstrates how this approach to normative modeling has been pursued by communities of users of the DDM.

## 3. A philosophical introduction to the DDM

We now turn to the target of our case study: the diffusion decision (or drift diffusion) model of decision making (DDM). The DDM is one member of a family of models that represent decision-making as a process of evidence accumulation, or adding up information over time. These models typically assume that decision-makers are forced to decide between two

[10] This is one of the first debates in the behavioral and brain sciences to turn a critical eye to notions of optimality shaping the collective scientific project. The debate concerned whether human economic decisions are "rational" according to formal theories developed in economics. Overviews of the debate can be found in Stanovich and West (2000) and Tetlock and Mellers (2002).

possible options (i.e., two-alternative forced decision-making), but variants that permit reasoning about more than two options continue to be developed (e.g., Tajima et al., 2019; Villarreal et al., 2024). 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).[11] 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 a wide range of models that can be construed at various spatiotemporal scales.

In what follows, we use Weisberg's (2013) notions of structure, scope, and assignment both to offer a philosophical introduction to the DDM. On Weisberg's account, a construal consists of a scope, assignment, and fidelity criteria; we use this final component to structure the case study in Section 4.

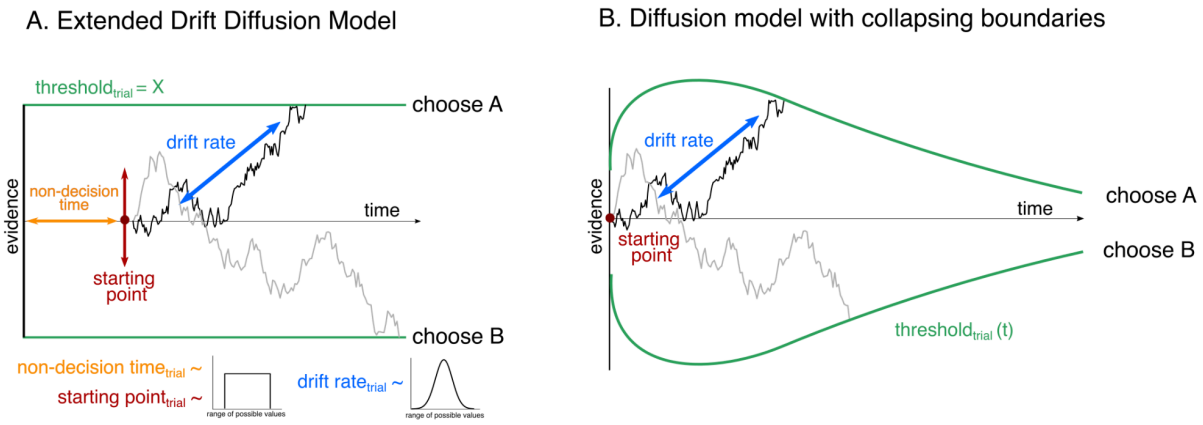


FIGURE 1: Graphical depictions of two standard forms of a diffusion model of speeded two-alternative choice. Both models assume that observers sequentially sample information, accumulate the difference in evidence between the two choice options, and commit to a

[11] 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).

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.

*3A: Formal structure of the DDM*

Recall that on Weisberg's (2013) account, the structure of a model refers to the mathematical objects and relations that comprise the formal representation of a target. In the DDM, this representation consists of components commonly termed starting point, decision variable, drift rate, internal noise, decision threshold, and non-decision time. This formal representation is presented graphically in Figure 1A. Mathematically, the evidence accumulation process operating in the DDM is expressed as:

$$dx = Adt + cdW, \quad x(0) = 0 \qquad (1)$$

where x(0) represents the starting point of the decision variable, *dx* represents a change in the decision variable *x* over a unit of time *dt*, *A* represents the drift rate, and *cdW* represents noise/diffusion in the accumulation process which follows a normal distribution with mean 0 and variance $c^2dt$ (Bogacz et al., 2006). This mathematical structure is equivalent to a random walk in probability theory or Brownian motion in physics. In the DDM, the structure commonly corresponds to a continuously-updating log likelihood ratio quantifying the probability that one of the two possible outcomes is correct, based on the evidence observed thus far. In other words, the decision variable reflects the time-evolving *difference* of evidence in favor of either option. The sampling/accumulation process ends when the value of the decision variable *x* exceeds a scalar threshold value *z*, at which point the decision maker commits to the choice corresponding to the threshold value reached (*z* or *-z*). The procedure whereby a sequential sampling model specifies how the accumulation process will be terminated and converted into a choice can be called the *decision rule* of the model.

This structure permits the DDM to generate predictions both of *which* option a decision-maker will choose and *how long* it takes them to commit to that choice on a trial-by-trial basis. This is what is meant when it is referred to as a joint model of choices and reaction times. Importantly, because of the noise term in the structure, the DDM assumes that

choice behavior is stochastic (i.e., subject to random variation). This structure permits the DDM to jointly predict a decision-maker's overall decision accuracy *and* the reaction time (RT) distributions corresponding to correct and incorrect decisions.

When the model is used to fit behavioral data, the starting point, drift rate, threshold, and non-decision time are commonly treated as "free parameters," meaning that the model fitting process aims to identify the values of these components that maximize the likelihood of the data being fitted. These can be contrasted with "model variables" which are components specified in the model whose values are either fixed by the user or vary probabilistically. Because the DDM formalizes the relationship between choices and RTs, the quantitative model fitting process is often tasked with fitting measurements from *both* of these elements, such that the model is really tasked with identifying parameter values that best describe how a decision maker traded off speed and accuracy. The resulting parameter values thus reflect the model's best estimate of the "settings" of the decision process that generated a particular pattern of target behavior. Based on the research question, it is often desirable to permit one or more the free parameters to vary as a function of experimental condition; this is what permits decomposing behavioral effects of experimental interventions into specific changes in the configuration of the latent decision process.

The structure we describe above corresponds to the "original" DDM (Stone, 1960). Presently, there are two variations of this original DDM that are most prominently used. Both of these forms retain the evidence accumulation process described in Equation 1, but posit different procedures within which that accumulation process generates choice behavior. The first process—commonly called the "extended DDM"—allows the drift rate, starting point, and non-decision time components to vary probabilistically on a trial-by-trial basis (Figure 1A). The second form—commonly called a "collapsing bound" diffusion model—posits that the threshold value for committing to a choice decreases over the course of a single decision (Figure 1B). This latter form initiated the "collapsing bound debate" that we conceptually analyze in Section 4.

*3B: Structure of the random dot motion task*

Much of the empirical success of the DDM is due to the development of the random dot motion discrimination task (Britten et al., 1992). 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). Importantly, although the overall *proportion* of coherently-moving elements remains constant within a trial, the actual elements whose position is displaced varies randomly with each refresh of the digital display. This "limited lifetime" property of the stimuli ensures that observers cannot make the decision via smooth visual pursuit: they must continuously sample the stimulus in order to accumulate evidence about the overall direction of coherence motion.

This task is useful for a number of reasons. First, it requires integrating evidence over time, so researchers have good reason to believe that the task utilizes the core mechanism of evidence accumulation (although see Stine et al. (2020) for challenges to this idea). Next, it is simple enough that rodent, primate, and human observers are all capable of performing it (Hanks & Summerfield, 2017), allowing for powerful cross-species comparisons. Finally, it permits a clean one-to-one mapping of stimulus properties onto components of the model, thus giving both experimenters and theorists alike a powerful tool for probing decision making processes under various tightly-controlled circumstances. The next subsection offers an overview of the wide-ranging pieces of empirical support for the DDM, the vast majority of which was generated using behavior in some variant of a motion discrimination task.

*3C: Assignment of components in the DDM*

Recall that, on Weisberg's (2013) account, a model's assignment specifies what each component of a model structure is meant to represent with respect to its target. It's important to note the standard assignment of components in the DDM that we describe here are specified with respect to the random dot motion task. Different applications of the DDM warrant slightly different assignments of model components, a point that can be overlooked in the meta-scientific writing that touts the utility of models for standardizing interpretations of behavior across task contexts. While it is true that the *structure* of the DDM provides common grounding for different measurements of behavior, how that structure is construed with respect to the target crucially depends on the measurements a user has of that target. For this reason, we highly encourage users to think critically about how components of the DDM are assigned to neural and/or cognitive processes as measured within a particular task setting (Bompas et al., 2023; Jones & Dzhafarov, 2014)

With those caveats in mind, we can now introduce the standard assignment of DDM components in the context of a motion discrimination task. 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. The starting point variable defines the initial value of the decision variable, and thus is conventionally interpreted as quantifying the bias an observer has toward one of the two choice options. The decision variable represents the decision maker's internal representation of accumulated evidence, sometimes called their time-evolving belief about the correct answer. The drift rate quantifies the rate of change in the decision variable per unit time, and thus is commonly thought to reflect the "quality" of the internal evidence driving a particular decision.[12] The non-decision time variable is intended to aggregate various kinds of delays in response times due to processing occurring before and after evidence accumulation. Processes believed to contribute to the non-decision time include

[12] More recently, the drift rate variable has also been construed to represent how quickly an observer internally processes information *in general* (Schubert et al., 2015*)*.

encoding of sensory information and initiating a motor response; in this sense, it can be thought of as a "convenience parameter," though some work has successfully decomposed this "non-"decision time into decision-relevant components (e.g. Kraemer & Gluth, 2023; Yoo & Bornstein, 2024). Finally, the internal noise variable is thought to reflect imperfections in the encoding and representation of sampled evidence, and also can be thought of as a "convenience variable" because its values are commonly fixed when the DDM is fit to human data (Ratcliff et al., 2016).

*3D: Empirical scope of the DDM*

On Weisberg's (2013) account, the scope of a model is the component of its construal that specifies which aspects of the target a user aims to capture in the model. Models thus have an intended empirical and/or theoretical scope and are initially evaluated with respect to those originally intended explananda. One of the reasons why the DDM is considered so successful is because it has continued to display explanatory adequacy for targets well outside its originally intended scope: identifying a formal relationship between choice accuracy and reaction time in human memory retrieval (Link & Heath, 1975; Ratcliff, 1978). Importantly, the DDM was developed as a purely cognitive model—nothing about the original construal mentioned neural activity or measurements as a desired component for the model to capture. It was not until the random dot motion task was developed by Britten et al. (1992) that the DDM was used *also* to reason about neural processes involved in two-alternative decision making.

Careful early studies conducted on non-human primates offered the first pieces of evidence for the DDM as a model of neural activity (Britten et al., 1996; Gold & Shadlen, 2001; Roitman & Shadlen, 2002; Shadlen & Newsome, 2001). These foundational studies provided evidence suggesting that both single-unit and population level recordings from distinct cortical regions exhibit activity that strongly resembles and correlates with distinct components of the DDM (see Gold & Shadlen (2007) for a review of this line of work). Since these early findings, the DDM has been used to link behavior with population-level neural responses in rodents (Brunton et al., 2013; Hanks et al., 2015; Khilkevich et al., 2024), as well as intracranial recordings, scalp oscillations, and blood oxygen level dependent signals in humans (Krueger et al., 2017; O'Connell & Kelly, 2021; Polanía et al., 2014; Weber et al., 2024). The early and growing evidence for the DDM as a model of the neural processes generating decision behavior has even led some researchers to posit evidence accumulation as a basic mechanism of decision making that is conserved across species (Hanks & Summerfield, 2017).

While this broad empirical scope lends a great deal of support to the DDM as a theory of speeded decision making, it can also result in confusion, ambiguity, and/or disagreement about precisely what inferences a user can make on the basis of applying the DDM to data. For this reason, among several others, it is desirable for modelers to explicitly discuss a model's intended construal when communicating their findings from the model.

## 4. Fidelity criteria driving model development and comparison in the DDM

In this section, we further demonstrate the utility of our philosophical tools for reasoning by using them to provide a novel conceptual analysis of the so-called "collapsing bounds" debate in the DDM. As shown in Figure 1, this debate concerns whether the threshold in the DDM ought to remain fixed over the course of a single decision (Figure 1A) or change as a function of time spent accumulating evidence (Figure 1B). Not only does each form of the model make highly accurate predictions about both behavioral and neural observations, but each form has also been proven, under different sets of assumptions, to optimize the two-alternative decision problem; much ink has thus been spilled about which form "correctly" represents the decision process (Evans et al., 2017; Hawkins et al., 2015; Miletić & van Maanen, 2019; Palestro et al., 2018; Ratcliff et al., 2016). Our review of the philosophy literature suggests that this question is misguided. This section aims to show that the two forms of the model are better thought of as *complementary* rather than *competing* tools for reasoning. Further, because the literature's rhetorical focus on competition has precluded consideration of when one form might be more suited for reasoning than the other, we tailored our conceptual analysis to addressing this gap.

Toward these ends, we first discuss two definitions of optimality employed by users of the DDM. In doing so, we demonstrate how formal definitions of optimality can be modified to increase their adequacy for particular purposes, and clarify the contexts in which each model optimizes the decision process. Then, we demonstrate how the two forms of the model reflect different explanatory aims among two subgroups of DDM users: cognitive psychometricians and theoretical decision (neuro)scientists. Whereas the latter group aims to develop normative theories of choice behavior, the former aims to develop models that permit statistically robust decomposition of behavior into model components.[13] These different aims warrant different styles of representational decision making and motivate different types of fidelity considerations. We conclude with an analysis of model comparison practices undertaken by users in each group, highlighting again how reasoning goals inform the the fidelity criteria determining a model's adequacy for purpose.

### *4A: Optimality criteria in the DDM*

Recall that the DDM establishes a formal link between decisions and the time it takes to make them; that is, it is a joint model of choices and reaction times. The speed-accuracy tradeoff in the DDM is determined by the threshold, the value of which specifies the "decision rule": at each point in time, determining whether an agent should keep sampling information or commit to a choice on the basis of already-accumulated evidence. Higher thresholds always result in increased choice accuracy, but often at the cost of taking longer to make a decision; lower thresholds allow subjects to make decisions more quickly but often come at the expense

[13] These two communities by no means exhaust the collection of DDM users, nor do we intend for the categories to be mutually exclusive; the distinction is simply useful for our narrative purpose.

of a greater number of errors. Optimality criteria in the DDM thus aim to define a formal benchmark for determining whether the threshold setting maximally balances speed and accuracy in a particular environment.

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). The definition of optimality that SPRT satisfies is that of *minimizing* the average amount of time (or samples) needed in order for the agent to respond at a predetermined level of accuracy (e.g., 80%). Crucially, the mathematical proofs of SPRT's optimality assume that the "difficulty" of the decision (e.g., the signal-to-noise ratio in the stimulus) is identical for all trials. In other words, the environment is *homogeneous* (Moran, 2015).

Because most experimental environments are actually heterogeneous (i.e., decision difficulty changes on each trial), it can be said that Bayes Risk lacks explanatory fidelity as an objective function for targets in these environments. A more common alternative in current work assumes that agents maximize their *reward rate*, defined as the 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 that extend beyond its heightened explanatory fidelity. For one, 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).

Formal analyses undertaken by Bogacz et al. (2006) and Moran (2015) offer helpful insights into the relationships between these optimality criteria and conditions under which each form of the DDM is optimal. The original DDM (i.e., without trial-variability in any parameters) is the continuous-time equivalent of SPRT, and thus is the optimal decision policy in homogeneous environments. Importantly, incorporating trial-level variability into any component of the original DDM not only breaks this normative status, but also makes evaluating the optimality of that form of the model mathematically impossible (Bogacz et al.,

2006). The collapsing bound DDM is the optimal policy in both heterogeneous environments (Drugowitsch, Moreno-Bote et al., 2012) and homogeneous environments, as in this latter case it simply collapses into the original DDM (Moran, 2015). Finally, Moran (2015) showed that Bayes Risk and reward rate optimality can be considered equivalent in the sense that decision thresholds which optimize reward rate also optimize Bayes Risk. Because collapsing bounds meet this criterion, Moran (2015) suggests they be considered "generally optimal" decision rules in the DDM.

*4B: Dynamical fidelity motivated the extended DDM*

Although the original DDM is commonly attributed to Ratcliff (1978), the model structure detailed in Ratcliff (1978) does not correspond to the structure that most authors refer to as the original DDM. Descriptions of the original DDM—where log-likelihoods are continuously estimated and decision boundaries are symmetric around a starting point of 0—can be found in Stone (1960) and Laming (1968). The structure of Ratcliff's (1978) model already reflects representational changes made to enhance the model's dynamical fidelity: allowing drift rates to vary on a trial-by-trial basis.[14]

Throughout every iteration of the DDM's development, Ratcliff's representational decisions were guided by the goal of creating a representation that can reproduce *all* measured aspects of human behavior on the two-alternative forced choice task (Ratcliff, 1978; Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002). Of uniquely high importance to Ratcliff is the ability of the DDM to reproduce patterns of RT distributions across task conditions, a point he emphasizes in nearly all papers involving the model (Ratcliff, 1978; Ratcliff et al., 2016; Ratcliff & McKoon, 2008). An empirically robust pattern in this regard is the mixture of slow and fast errors present in a sample of measurements. This mixture varies both across individuals performing a task with the same instructions, and within a single individual when they are instructed to prioritize either the speed or accuracy of their responses. Ratcliff's solution to capturing this empirical property of his target was to allow the drift rate and the starting point of the evidence accumulation process to vary probabilistically from trial-to-trial (Ratcliff & Rouder, 1998). A later modification—trial-wise variability in the non-decision term—was added to increase the DDM's ability to fit RT distributions with a high amount of variance in the 0.1 quantile (Ratcliff & Tuerlinckx, 2002). The extended DDM encompasses all of these changes and functions as the standard form of the model in current research (Ratcliff et al., 2016).

As mentioned in the previous section, the representational changes that made the DDM satisfy Ratcliff's dynamical fidelity criteria also break the normative status of its original form. But because its structure still posits a *procedure* for the evidence accumulation process and permits decomposition of the target's behavior into components of that procedure, the model continues to give a mechanistic explanation. Importantly, the evidence accumulation *process* in

[14] We thank Barbara Dosher for directing our attention to this detail.

the DDM (i.e., the mathematical form of the decision variable) is the normative solution to two-alternative forced decisions in any environment (Moran, 2015). It is thus the details about *how* that process is converted into decisions (i.e., incorporating trial-wise variability) that break the optimality of the procedure as a whole. If the extended DDM is tasked with answering the question of *why* choice behavior exhibits the patterns that it does, it does so by appealing to an optimal evidence accumulation process that is stochastically translated into choices.

Crucially, the *structure* of this stochasticity in the DDM was determined on the basis of statistical best fit to a broad range of data. We argue that this structure reflects the aims of cognitive psychometrics as a whole: developing principled models of cognitive processes that permit robust and reliable decomposition of behavior into meaningfully different component parts. Researchers in this subcommunity of DDM users often use statistical parsimony to guide their formal theory building, a position that naturally lends itself to structures with components that prioritize compact description over normative guarantees (Palminteri et al., 2017; Vandekerckhove et al., 2015). In this sense, statistical parsimony is a primary explanatory fidelity criterion for users with these goals. Because the extended DDM satisfies both the high standards of dynamical fidelity and a guiding explanatory fidelity criterion, it can be considered to have been "optimized" for the explanatory goals of cognitive psychometrics.

*4C: Explanatory fidelity motivated the collapsing bound diffusion model*

The structure of the collapsing bound diffusion model, we argue, has likewise been "optimized" to meet the explanatory goals of users we call theoretical decision (neuro)scientists. This community is considerably more heterogeneous than the one discussed above, comprised both of users whose focus is on behavior (Frazier & Yu, 2007) and those who use the DDM to link behavior with neural activity (e.g., Drugowitsch, Moreno-Bote et al., 2012). The feature uniting these users is their commitment to optimality as an explanatory fidelity criterion. This section thus discusses how the collapsing bound diffusion model is the result of representational decisions that aim to preserve the normative status of the model under assumptions that are more aligned with the contexts in which the target's behavior is measured.

The first specification of a diffusion model with collapsing decision boundaries was proposed by Frazier and Yu (2007). These authors identified the following assumption required for the optimality of the original DDM: that agents have an unlimited amount of time to perform the evidence accumulation process (formally called the assumption of infinite horizon). Because most experimental tasks place a hard limit on the amount of time an agent can spend on each trial, this formal representation of the task environment is poorly aligned with the environment in which the target's behavior is measured. Frazier and Yu (2007) propose a different normative model that incorporates stochastic (i.e., randomly varying) time limits and found that the optimal procedure in this specification involves decision thresholds that decrease over the course of a single choice. In a similar vein, Drugowitsch, Moreno-Bote et al. (2012) set out to find a model structure specifying the optimal procedure for making decisions

in a heterogeneous task environment, i.e., one where decision difficulty (or signal-to-noise ratio) varies on each trial. In order to do this, they introduced a new component to the normative formal representation: a small cost for each piece of sampled evidence. The optimal procedure on this representation also involves thresholds that decrease over the course of a decision.

In order to identify normative solutions to these differently-formalized problems, Frazier and Yu (2007) and Drugowitsch, Moreno-Bote et al (2012) both made use of a criterion known as Bellman optimality. This criterion posits that an optimal procedure is one that maximizes an agent's expected return, i.e., the amount of reward earned over the course of an experiment. Contrary to the mathematical proofs underpinning the original DDM's optimality (Bogacz et al., 2006; Edwards, 1965), identifying Bellman-optimal solutions requires using a computational procedure known as dynamic programming. Details of the procedure vary across applications, but the general idea is that agents recursively update an estimate of their expected return with each new piece of information they get from the environment (Sutton & Barto, 2018). Crucially, this means that the precise form of the Bellman-optimal procedure will depend on details of the environment (e.g., strength of evidence on each trial, reward scheme, timing, etc.), and thus will vary across tasks. But if the mathematical form is determined using a Bellman equation, users have a guarantee that the form maximizes expected return in that environment. In the case of the DDM, collapsing bounds thus maximize expected return for evidence accumulation decisions made both within a fixed amount of time and in heterogeneous environments.

As we have emphasized in our toolkit, it is crucial that model users pay close attention to the details of a model's structure and construal when making statistical inferences about the target on the basis of the model. One instructive example is offered by Drugowitsch, Moreno-Bote et al. (2012), who helpfully point out that when the assumption of a homogeneous environment is relaxed, the structure of the decision variable (Equation 1) can no longer be construed as the log-odds of either choice being correct. In order for the structure to represent this quantity, decision makers must know exactly what the strength of evidence is on each trial, which—barring specific task manipulations to give agents that information—is not the case in heterogeneous environments (Drugowitsch, Moreno-Bote et al., 2012, p. 3622). This point is subtle but crucial for construing models that posit a diffusion process as the evidence accumulation mechanism, especially when they are used to infer the coding properties of particular neurons and/or brain regions (e.g., Rorie et al., 2010).

A procedure whereby evidence accumulates to a threshold value that decreases over time thus emerges as a generally normative solution to slightly less idealized formalizations of the speeded two-alternative decision problem. This structure reflects the explanatory aims of DDM users who desire normative explanations for choice behavior. When the assumptions required for the original DDM's normativity were too restrictive, users in this community opted to identify optimal structures on weaker assumptions (i.e., finite amount of time to decide and different stimulus difficulty on each trial). This approach preserves the model's normative status while also enhancing its fidelity with respect to explanatory aims of its users.

*4D: A principled heuristic for choosing between the forms*

The last two sections have demonstrated how different reasoning goals—operationalized as fidelity criteria—led to the development of the two standard forms of the DDM. In contrast to some of the rhetoric of previous model comparisons surveyed below, we argue that these two forms offer complementary perspectives on the mechanisms of decision making. Our suggestion is that the appropriate form to use depends on your explanatory goals. If you wish to offer (or test) a *normative* explanation for speeded two-alternative decisions in heterogeneous and/or time-limited environments, then you must use the collapsing bound model. But if the normativity of the process is irrelevant for your explanatory or reasoning goals, then the extended DDM is fully adequate for that purpose. Of course, it is always possible to directly compare how well each form explains the behavior of the target on a particular task. But as we show in the next section, performing principled comparisons between the forms is far from trivial.

*4E: Dynamical and explanatory fidelity in model comparison*

Comparing the performance of different models with respect to fidelity criteria of interest is the primary means whereby model users determine which model structure offers the best explanation for their target. This final section contrasts two approaches—statistical and formal—that researchers have taken to compare the extended DDM to collapsing bounds diffusion models.

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 utilized three different quantitative metrics of model performance—Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and nested likelihood ratio tests. The authors reported that all three metrics returned similar results, which was their justification for reporting only BIC values in the published manuscript. The BIC is a heuristic metric that aims to approximate Bayes Factors, which are performance criteria that quantify the relative evidence in support of each model based on the empirical data it is tasked with fitting (Rouder & Morey, 2012). Because Bayes Factors can only be directly computed for models fitted with Bayesian methods, the BIC offers an approximation that can be estimated for models fitted with any statistical method.

Importantly, the BIC tends to overpenalize model complexity, which is commonly defined as the number of free parameters in the model[15]. We argue that this metric, although appropriately calibrated for the fidelity criteria of cognitive psychometricians characterized in

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

Section 4B, put the collapsing bounds models at a quantitative disadvantage. All of the collapsing bounds models tested by Hawkins et al. (2015) were specified in a manner that made them between 1.3 to 2 times as complex as the extended DDM. One of sources of this heightened complexity was the incorporation of trial-level variability in drift rate, starting point, and non-decision time into the collapsing bounds models that did not originally incorporate this variability (Drugowitsch, Moreno-Bote et al., 2012; Frazier & Yu, 2007). Hawkins et al.'s (2015) logic behind this decision was to create nested model structures, such that the extended DDM represents the simplest possible decision process in the comparison. This type of comparison set has been called a "stacked deck," since it quantitatively favors the null model (i.e., the extended DDM). However, 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 collapsing bounds models. These changes better aligned their specifications of collapsing bounds with the fidelity criteria motivating this form of the model, since neither Drugowitsch, Moreno-Bote et al. (2012) nor Frazier and Yu (2007) needed to incorporate any trial-varying parameters to achieve their fidelity criteria. 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 crucial 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 quantitative metrics assess *relative* measures of performance, it is imperative that alternative models are specified in a way that permit principled and meaningful contrasts. Ideally, these specifications consider ***both*** the structure and construal of alternative models for maximally informative comparisons. Finally, 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, as their structure can make all the difference with respect to the evidence of a "primary" model's performance.

We can contrast Hawkins et al.'s (2015) statistical approach with the formal approach Moran (2015) uses to answer two questions. First, Moran (2015) investigates the formal relationship between the two optimality criteria discussed in Section 4A: Bayes Risk and reward rate. Moran (2015) finds that they are functionally equivalent criteria of optimality, such that any threshold that optimizes one criterion also optimizes the other. This is a notable insight because previous discussions of optimality in the DDM (e.g., Bogacz et al., 2006) have instead emphasized how these optimality criteria differ from each other. By formalizing their conceptual equivalence, Moran's (2015) finding permits newer ways of thinking about normative forms of the DDM. Next, Moran (2015) seeks a normative solution for making decisions in environments

that are not only heterogeneous, but also *biased*, i.e., one of the choice outcomes is more commonly correct or rewarded than the other. To do this, Moran conducts an analysis that showcases the utility of formal reasoning: identifying the optimal solution ***given*** that the underlying procedure corresponds to the original DDM. This approach to normative modeling differs from those surveyed in Section 4C in that Moran (2015) seeks to identify a form of the model that maximizes reward rate using a process known to be *sub*optimal, licensing inferences about the optimality of a technically suboptimal solution.

Contrasting the results of Moran's (2015) analysis with those of van Ravenzaaij et al. (2012) further demonstrates the utility of formal approaches to model comparison. On the basis of several simulations, van Ravenzwaaij et al. (2012) reported that maximizing reward rate in biased and heterogeneous environments requires adjusting only the starting point of the decision process. This finding contradicted previous theoretical work indicating that both the starting point and drift rate must be biased (Bogacz et al., 2006), as well as empirical evidence supporting the existing of a biased drift in human and non-human primate decision making (Hanks et al., 2011). In approaching the same question from a different perspective, Moran (2015) identified a key oversight in van Ravenzwaaij et al.'s (2012) simulations: all of their models used the same threshold value, which was chosen arbitrarily. This representational decision simplified the space of possible solutions in a manner that aligned with van Ravenzwaaij et al.'s (2012) specific goal (i.e., challenging the model reported in Hanks et al., 2011). However, it also omitted the fact that optimal solutions *also* depend on how/where the threshold value is set. Moran (2015) then showed that when the same simulations reported by van Ravenzwaaij et al. (2012) included threshold values in the search space (along with negative values of drift rate), the reward-rate maximizing process is one that imparts a bias *both* on the starting point and drift rate.

In sum, these contrastive examples highlight how representational decisions motivated by particular reasoning goals fundamentally shape findings in model-based research. Further, they demonstrate that there are a variety of approaches users can take when comparing models. While this diversity promotes creativity in analytic reasoning, it can also breed confusion and/or disagreement about the validity of a particular approach or set of results. Our goal in this analysis—and the toolkit that preceded it—was to offer a novel conceptual perspective on this fundamental challenge in model-based research.

## 5. Summary and Conclusion

In this article, we summarized some core ideas in philosophy of modeling to develop a "philosophical toolkit" for computational cognitive modeling in Section 2. We then demonstrated the utility of such a resource by using it to give a philosophical introduction to an extremely prominent model in the brain and behavioral sciences (Section 3) and then offer a novel conceptual analysis of a long-standing debate regarding the form of that model (Section 4). Throughout, we emphasized the central role that *reasoning goals* play in shaping every step

of model-based research, echoing recent calls from philosophy (Danks, 2015; Potochnik & Sanches de Oliveira, 2020), computational neuroscience (Kording et al., 2018), and cognitive modeling of human behavior (Wilson & Collins, 2019). Our goal in doing so was to highlight the user-dependence of the insights afforded by these formal tools, an aspect that can be overlooked by both new and seasoned researchers alike. This goal was motivated by two core contributions that philosophy can offer practicing scientists: (1) highlighting implicit beliefs or assumptions they have about how their tools work, and (2) identifying the logical and epistemic limitations of those tools with respect to particular goals. Our toolkit and case study, though focused on topics in the DDM, were constructed with the goal of conveying insights that can generalize to most, if not all, of the formal models that are becoming increasingly common in cognitive neuroscience.

**References**

Ambekar, A., Ward, C., Mohammed, J., Male, S., & Skiena, S. (2009, June 28). Name-ethnicity classification from open sources. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD09: The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris France. https://doi.org/10.1145/1557019.1557032

Anderson, J. R. (1990). *The adaptive character of thought*. Lawrence Erlbaum Associates.

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

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

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.

Bompas, A., Sumner, P., & Hedge, C. (2023). Non-decision time: the Higg's boson of decision. In *bioRxiv* (p. 2023.02.20.529290). https://doi.org/10.1101/2023.02.20.529290

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

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.

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

Chintalapati, R., Laohaprapanon, S., & Sood, G. (2018). Predicting race and ethnicity from the sequence of characters in a name. In *arXiv [stat.AP]*. arXiv. http://arxiv.org/abs/1805.02109

Colombo, M., & Knauff, M. (2020). Editors' review and introduction: Levels of explanation in cognitive science: From molecules to culture. *Topics in Cognitive Science*, *12*(4), 1224–1240.

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*.

Clarendon Press.
Danks, D. (2015). Goal-dependence in (scientific) ontology. *Synthese*, *192*(11), 3601–3616.
Dayan, P., & Abbott, L. F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.
Drayson, Z. (2020). Why I am not a literalist. *Mind & Language*, *35*(5), 661–670.
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.
Dworkin, J. D., Linn, K. A., Teich, E. G., Zurn, P., Shinohara, R. T., & Bassett, D. S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience*, *23*(8), 918–926.
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.
Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports*, *7*(1), 16433.
Figdor, C. (2018). *Pieces of mind: The proper domain of psychological predicates*. Oxford University Press.
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., 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/
Frigg, R., & Nguyen, J. (2017). Models and representation. In *Springer Handbook of Model-Based Science* (pp. 49–102). Springer International Publishing.
Gamboa, J. P. (2024). On cognitive modeling and other minds. *Philosophy of Science*, *91*(3), 615–633.
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.

Hanks, T. D., & Summerfield, C. (2017). Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, *93*(1), 15–31.

Harhen, N. C., & Bornstein, A. M. (2023). Overharvesting in human patch foraging reflects rational structure learning and adaptive planning. *Proceedings of the National Academy of Sciences*, *120*(13). https://doi.org/10.1073/pnas.2216524120

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.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Jones, M., & Dzhafarov, E. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychology Review*, *121*(1), 1–32.

Kaplan, D. M. (2017). *Explanation and integration in mind and brain science*. Oxford University Press.

Khilkevich, A., Lohse, M., Low, R., Orsolic, I., Bozic, T., Windmill, P., & Mrsic-Flogel, T. D. (2024). Brain-wide dynamics linking sensation to action during decision-making. *Nature*, 1–11.

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

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

Kraemer, P. M., & Gluth, S. (2023). Episodic Memory Retrieval Affects the Onset and Dynamics of Evidence Accumulation during Value-based Decisions. *Journal of Cognitive Neuroscience*, *35*(4), 692–714.

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.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*(2), 279–311.

Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *The Behavioral and Brain Sciences*, *43*, e1.

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

Lin, Y.-S., & Strickland, L. (2020). Evidence accumulation models with R: A practical guide to hierarchical Bayesian methods. *The Quantitative Methods for Psychology*, *16*(2), 133–153.

Miletić, S., & van Maanen, L. (2019). Caution in decision-making under time pressure is mediated by timing ability. *Cognitive Psychology*, *110*, 16–29.

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

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

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

Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychonomic Bulletin & Review*, *25*(4), 1225–1248.

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.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433.

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.

Potochnik, A. (2018). *Idealization and the aims of science*. University of Chicago Press.

Potochnik, A., & Sanches de Oliveira, G. (2020). Patterns in Cognitive Phenomena and Pluralism of Explanatory Styles. *Topics in Cognitive Science*, *12*(4), 1306–1320.

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.

Rorie, A. E., Gao, J., McClelland, J. L., & Newsome, W. T. (2010). Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. *PloS One*, *5*(2), e9308.

Ross, L., & Woodward, J. (2023). Causal Approaches to Scientific Explanation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903.

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., & Bergmann, K. (2015). Decomposing the relationship between mental speed and mental abilities. *Intelligence*, *51*, 28–46.

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.

Shinn, M., Lam, N. H., & Murray, J. D. (2020). A flexible framework for simulating and fitting generalized drift-diffusion models. *eLife*, *9*. https://doi.org/10.7554/eLife.56938

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*(5), 645–665; discussion 665–726.

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between

integration and non-integration strategies in perceptual decision making. *eLife*, *9*. https://doi.org/10.7554/eLife.55365

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

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction, 2nd edition*. MIT Press.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*, 449–508.

Tajima, S., Drugowitsch, J., Patel, N., & Pouget, A. (2019). Optimal policy for multi-alternative decisions. *Nature Neuroscience*, *22*(9), 1503–1511.

Tetlock, P. E., & Mellers, B. A. (2002). The great rationality debate. *Psychological Science*, *13*(1), 94–99.

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint Models of Neural and Behavioral Data* (1st ed.) [PDF]. Springer Nature.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. (2015). Model Comparison and the Principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 300–319). Oxford University Press.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, *1*(3), 127–128.

Villarreal, M., Chávez De la Peña, A. F., Mistry, P. K., Menon, V., Vandekerckhove, J., & Lee, M. D. (2024). Bayesian graphical modeling with the circular drift diffusion model. *Computational Brain & Behavior*, *7*(2), 181–194.

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

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22.

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.

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.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*. https://doi.org/10.7554/eLife.49547

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.

Yoo, J., & Bornstein, A. (2024). Temporal dynamics of model-based control reveal arbitration between multiple task representations. In *PsyArXiv*. https://doi.org/10.31234/osf.io/sgcy5

Zhou, D., Bertolero, M. A., Stiso, J., Cornblath, E. J., Teich, E. G., Blevins, A. S., Oudyk, K., Michael, C., Urai, A., Matelsky, J., Virtualmario, Camp, C., Castillo, R. A., Saxe, R., Dworkin, J. D., & Bassett, D. S. (2022). *Diversity Statement and Code Notebook v1.1.2*. Zenodo. https://github.com/dalejn/cleanBib