Mutual benefits: Combining reinforcement learning with sequential sampling models

Miletić, S.; Boag, R.J.; Forstmann, B.U.

DOI
10.1016/j.neuropsychologia.2019.107261

Publication date
2020

Document Version
Final published version

Published in
Neuropsychologia

Neuropsychologia 136 (2020) 107261

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

Mutual benefits: Combining reinforcement learning with sequential sampling models

Steven Miletić [*,1], Russell J. Boag [1], Birte U. Forstmann

University of Amsterdam, Department of Psychology, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Sequential sampling models
Reinforcement learning
Instrumental learning
Decision-making

ABSTRACT

Reinforcement learning models of error-driven learning and sequential-sampling models of decision making have provided significant insight into the neural basis of a variety of cognitive processes. Until recently, model-based cognitive neuroscience research using both frameworks has evolved separately and independently. Recent efforts have illustrated the complementary nature of both modelling traditions and showed how they can be integrated into a unified theoretical framework, explaining trial-by-trial dependencies in choice behavior as well as response time distributions. Here, we review a theoretical background of integrating the two classes of models, and review recent empirical efforts towards this goal. We furthermore argue that the integration of both modelling traditions provides mutual benefits for both fields, and highlight promises of this approach for cognitive modelling and model-based cognitive neuroscience.

1. Introduction

The field of model-based cognitive neuroscience uses cognitive models to bridge the gap between neuroimaging data and behavior (Forstmann et al., 2011; Forstmann and Wagenmakers, 2015; Gläscher and O'Doherty, 2010; O'Doherty et al., 2007; Turner et al., 2019b, 2017a). Cognitive models constitute a cognitive theory and a corresponding psychometric model simultaneously, by decomposing empirically observed behavior into the latent cognitive processes that are thought to have caused the behavior. By placing such models "in the middle" between behavioral data and neural data, they provide a unified framework through which to interpret behavioral and neural data (Fig. 1).

This approach has provided significant progress in our understanding of the neural underpinnings of cognition, especially using two distinct classes of cognitive models: Reinforcement learning (RL) models on the one hand, and sequential sampling models (SSMs) of decision making on the other. Hallmark studies in RL (for reviews, Gershman and Daw, 2017; O'Doherty et al., 2017) used reinforcement learning models to identify brain regions that encode (experienced and predicted) value (e.g., Daw et al., 2006; O'Doherty et al., 2003; see also O'Doherty, 2014), neural correlates of reward prediction errors (e.g., O'Doherty et al., 2003; Schultz et al., 1997), and the neural underpinnings of different learning systems (e.g., Daw et al., 2011; Gläscher et al., 2010). Model-based cognitive neuroscience using SSMs (for reviews, Forstmann et al., 2016; Mulder et al., 2014) shed light on the neural mechanisms underlying response caution (Forstmann et al., 2010, 2008; Van Maanen et al., 2011), choice biases and perceptual biases (Leong et al., 2019; Mulder et al., 2012), attention effects (Nunez et al., 2017, 2015), and provided insight into the neural basis of evidence accumulation (Gold and Shadlen, 2001; Purcell et al., 2010; Shadlen and Newsome, 2001).

Recent papers (Fontanesi et al., 2019a, 2019b; Frank et al., 2015; Millner et al., 2018; Pedersen et al., 2017; Sewell et al., 2019) proposed that the two modelling classes can be unified into a single theoretical framework. On first sight, such a merger seems at odds with the fact that these modelling frameworks appear different in both the behavior they seek to explain, as well as the mechanisms proposed to do so. For RL models (Sutton and Barto, 2018), the target behavior that is to be explained, is changes in choice behavior due to error-driven learning in value-based decision making. Whereas implementations differ between RL models, the core of the explanation of this change in behavior is that

* Corresponding author.
E-mail address: s.miletic@uva.nl (S. Miletić).
[1] Authors contributed equally.

https://doi.org/10.1016/j.neuropsychologia.2019.107261
Received 20 August 2019; Received in revised form 21 October 2019; Accepted 10 November 2019
Available online 14 November 2019

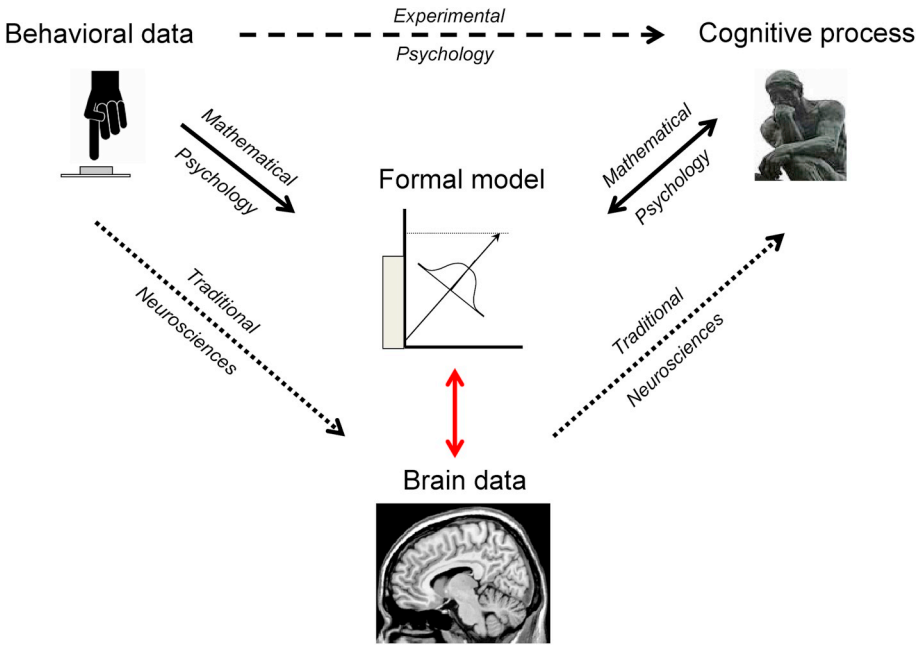

Fig. 1. Illustration of the "model in the middle" approach. Traditional experimental psychology studies cognitive processes using behavioral data; traditional cognitive neuroscience studies cognitive processes using neural data combined with constraints from behavioral data. Mathematical psychology studies cognitive processes by formally modelling behavioral data, which are used to inform analyses of neural data in the "model in the middle" approach. Adapted from Forstmann et al. (2011) with permission.

people maintain representations of expected value (or utility), inform their choices based on these expected values, and subsequently update them based on the difference between the expected value and actual outcome – which is referred to as the *reward prediction error*.

RL models using error-driven learning account for changes in the relative choice probabilities over the course of learning, but do not typically account for concurrent (changes in) response latencies, a dimension of the data that is often entirely ignored in the employed cognitive models. In contrast, SSMs (Forstmann et al., 2016; Ratcliff et al., 2016) provide a comprehensive mechanistic account of the decision stage, simultaneously explaining how choice accuracy and latencies arise from a common set of latent cognitive processes (e.g., information processing speed, response caution, motor response time). Like in the RL domain, specific implementations differ between SSMs, but they share the general idea that people accumulate evidence over time for each choice option, until they reach a threshold level of evidence, at which point they commit to a decision. Although practice effects have been of considerable interest in perceptual decision-making (Dutilh et al., 2009; Evans et al., 2018; Liu and Watanabe, 2012; Petrov et al., 2011) in most applications, SSMs do not provide a mechanistic account of how the processes driving decision making (e.g., processing speed, response caution) might be changed or adjusted as learning progresses (e.g., via an explicit learning mechanism).

In short, the RL framework explains learning over time, but provides no mechanistic account of decision making and ignores response time distributions. The SSM framework explains how decisions arise from an evidence accumulation process and provide detailed fits to response time distributions, but often ignores learning over time. From this perspective, the RL and SSM frameworks are complementary. In what follows, we first review a theoretical grounding for the integration of the two classes of models, followed by a review of the empirical efforts towards combining RL models with SSMs. Afterwards, we argue that integration can work to the benefit of both modelling traditions: Accounting for response time distributions provides a major benefit for RL models, and the inclusion of learning a major benefit for SSMs. Finally, we highlight the exciting promises for cognitive science as well as model-based cognitive neuroscience.

2. Linking RL and SSMs: Theory

To illustrate the general approach, we assume benchmark models for both the (so-called model-free[2]) RL and SSM frameworks, and assume the learning task to which the models are applied consists of two stimuli. As briefly noted above, the core assumption of the RL models is that agents maintain an internal representation of the expected value (or utility) associated with each stimulus and/or action, translate these expected values into choices, and update their expected values based on the mismatch between the expected values and actual outcomes. To formalize these notions, RL models consist of at least two components: An update rule and a choice rule (or policy). Update rules define prediction errors and specify how subjective reward expectations change as a function of these prediction errors. They are based on the temporal difference (Rescorla and Wagner, 1972) between the actual and expected outcomes, such as the so-called *state-action-reward-state-action* (SARSA) rule:

$$Q_{i,t+1} = Q_{i,t} + \alpha\left(r_t - Q_{i,t}\right) \quad 1$$

where Q is the expected value for choice option i on trial t, and r is the reward. The difference between the reward and the Q-value is defined as the prediction error. Finally, α is a free parameter called the *learning rate*. Higher learning rates indicate that Q-values fluctuate heavily, and consequently, each choice is mostly influenced by the rewards received on a small subset of previous trials. Conversely, lower learning rates indicate that Q-values evolve slowly, and choices are based on a larger subset of previous trials. For simplicity (but without loss of generality), we ignore the many extensions to this basic rule, including multiple learning rates for negative and positive prediction errors (Christakou et al., 2013; Daw et al., 2002; Frank et al., 2009; Gershman, 2015; Haughey et al., 2007; Niv et al., 2012), eligibility traces to allow for

[2] Note that the terms model-free and model-based reinforcement learning refer to the absence or presence of an internal model of the outside world in an agent's mind. Despite the similarity in nomenclature, these concepts are distinct from model-free and model-based cognitive neuroscience.

updating of previously visited states (Bogacz et al., 2007), choice perseveration to model the tendency to repeat previous choices (Christakou et al., 2013; Worthy et al., 2013; Yechiam and Ert, 2007), and mixture models of model-free and model-based learning (Daw et al., 2011; Daw and Dayan, 2014; Dezfouli and Balleine, 2013; Doll et al., 2015; Kool et al., 2018).

The second component of RL models is the choice rule, which is classically called the soft-max function. Soft-max specifies the probability of choosing choice option i (out of N options) as:

$$p(i) = \frac{e^{\beta Q_i}}{\sum_j^N e^{\beta Q_j}} \quad 2$$

where β is a free parameter known as the inverse temperature, and Q_i is the expected value for choice option i. With two choice alternatives, say A and B, this can be rewritten as:

$$p(B) = \frac{e^{\beta Q_B}}{e^{\beta Q_A} + e^{\beta Q_B}} = \frac{1}{1 + e^{\beta (Q_A - Q_B)}} \quad 3$$

highlighting that, in a two choice-alternative task, the *difference* in Q-values for both choice options drives choice probabilities.

Crucially, soft-max only makes predictions for choice probabilities, and ignores a second dimension of behavioral data: response latencies. This is not to say that response times are not acknowledged as a valuable source of information in value-based decision making (Krajbich et al., 2015). In fact, SSMs, which exploit information in response time, have been applied very successfully in the domain of value-based decision-making as well (Busemeyer et al., 2019; Krajbich et al., 2010; Krajbich and Rangel, 2011; Milosavljevic et al., 2010; Polanía et al., 2014; Rodriguez et al., 2015, 2014; Turner et al., 2018b). Note, however, that contrary to the literature reviewed below, these papers did not explicitly model learning dynamics.

The fundamental assumption of sequential sampling models is that, in order to decide, participants accumulate noisy evidence until a threshold level of evidence is reached, and a decision is made. At that point, the participant initiates a motor response. Fig. 2 provides a schematic overview of the decision process as proposed by the most popular SSM, the diffusion decision model (DDM; Ratcliff, 1978; Ratcliff et al., 2016). The DDM specifically assumes that the difference in evidence for both choice options is accumulated (i.e., the net 'attractiveness' of both choice options), according to:

$$\frac{dx}{dt} = vdt + sdW \quad 4$$

where x is the amount of accumulated evidence, t is time, v is the mean speed of evidence accumulation (the *drift rate*), and dW is Gaussian noise with standard deviation s (the *diffusion coefficient*). In words, on each time step after stimulus onset, a sample is drawn from a Gaussian distribution with mean v and standard deviation s, which are accumulated until either threshold $-A$ or A is reached. Reaching a threshold amounts to committing to a decision, at which point the motor processes that allow for an overt response (typically a button press) are initiated. Evidence accumulation is often thought to start midway between the two response thresholds, but the start point parameter z can be shifted towards either choice boundary to incorporate a bias in the prior beliefs about the two choice options. The resulting response time is the sum of the time of evidence accumulation (the decision time), plus an intercept to account for the time it takes to perform initial perceptual processes and motor execution, which are collectively known as the non-decision time, $t0$. For simplicity, extensions to this model (such as between-trial variability in drift rate, start point, and non-decision time) are ignored here.

2.1. A combined RL-DDM

The DDM explains choice and response latencies in terms of an accumulation-to-bound process parametrized by v, A, z and $t0$, whereas traditional reinforcement learning models use soft-max as a choice function only. A logical departure point to merge the RL models and SSMs is therefore to substitute soft-max with the DDM, or vice versa, augment the DDM with an update rule such as SARSA. A requirement to do this is a *linking* function: A mapping between aspects of the learning model (parameters or internal dynamics such as expected values or prediction errors) and DDM parameters. One intuitive linking function follows from a comparison between soft-max and the DDM choice function, which is given by (parametrisation from Bogacz et al., 2006):

$$p_{error} = \frac{1}{1 + e^{2vA/s^2}} - \frac{1 - e^{-2zv/s^2}}{e^{2vA/s^2} - e^{-2vA/s^2}} \quad 5$$

Assuming an unbiased diffusion process (i.e., $z = 0$), the probability of an error simplifies to:

$$p_{error} = \frac{1}{1 + e^{2vA/s^2}} \quad 6$$

If we further assume that the drift rate equals the difference between the evidence for both choice options ($v = Q_A - Q_B$), we obtain

$$p_{error} = \frac{1}{1 + e^{(Q_A - Q_B) * 2A/s^2}} \quad 7$$

which is formally equivalent to a soft-max function with the inverse temperature parameter equal to the ratio $2A/s^2$ Conceptually, this is the amount of evidence required to commit to a decision, relative to the amount of noise in evidence accumulation. In other words, the DDM with the drift rate parameter defined as $v = Q_A - Q_B$ provides an identical *choice* function as soft-max, but adds a *latency* function, of which the exact shape depends on the ratio of the additional parameters A and s. In principle, the DDM is therefore able to fit the exact same choice patterns as soft-max (which is important because soft-max has been shown to fit well to empirical data) while adding a prediction of entire response time distributions.

Note that this equivalence has been described before (Tuerlinckx and De Boeck, 2005) and a linear mapping between value differences and drift rate has been used to fit the DDM to empirical value-based decision data (Fontanesi et al., 2019a, 2019b; Millner et al., 2018; Milosavljevic et al., 2010; Pedersen et al., 2017). Note further that the ability to mimic soft-max is not unique to the DDM, as other SSMs are able to either approximate soft-max or can even be formally equivalent as well (notably, Tuerlinckx and De Boeck, 2005, describe a second equivalence using a racing accumulator model).

We highlight the mathematical relations between soft-max and SSMs here for three reasons. Firstly, they offer a natural departure point to link SSM parameters with variables from RL models, while acknowledging that it is an empirical question whether this linking function provides the best account of behavioral data. Secondly, these relations provide an interesting cognitive interpretation of the soft-max inverse temperature parameter in terms of response caution (i.e., thresholds), which we discuss further below. Thirdly, it shows that SSMs such as the DDM form a generalization of soft-max into the time domain. Speculatively, the SSM may approximate the actual cognitive processes underlying value-based decision making in learning tasks, and soft-max captures the choice function because it is an instantiation (and simplification) of the SSM choice rule.

The mathematical relations between SSMs and soft-max are only theoretical in nature, and empirical testing is required to assess whether SSMs can in fact describe choice behavior in learning tasks. As

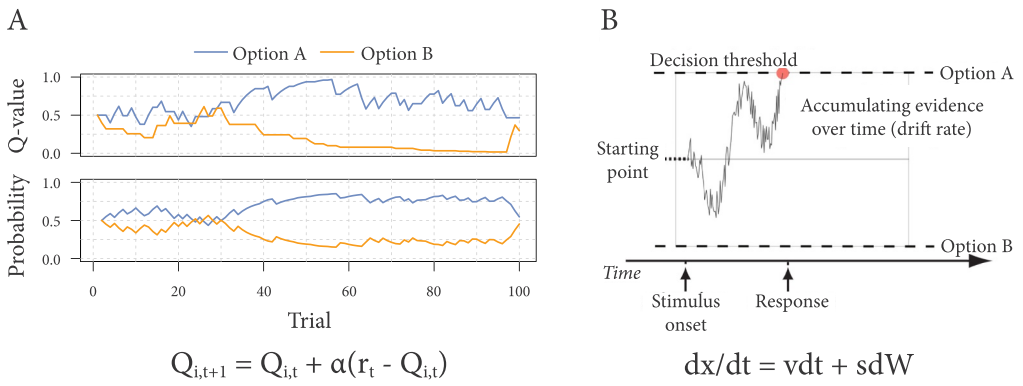

Fig. 2. A) Evolution of Q-values over time due to learning (upper panel) and the corresponding choice probabilities (lower panel) on a single-trial basis. In traditional RL modelling, the choice probabilities are determined by the soft-max function. With RL-SSMs, these choice probabilities are a function of the SSM (e.g., the DDM). B) Illustration of decision processes as formalized in the diffusion decision model (DDM). To reach a decision, participants accumulate noisy evidence for both choice options (A and B) until one of the two thresholds is reached and a decision is made. The choice corresponds to the threshold that is reached, and the response time is the time it took to reach this boundary, plus the time it takes for perceptual encoding and motor response.

preliminary evidence, it is possible to derive some simple predictions about empirical data in learning tasks, by reasoning about the cognitive processes proposed by SSMs such as the DDM. More specifically, if we assume that the DDM is the choice function underlying decisions in an instrumental learning task, and the choice data is modelled using soft-max, then we can derive several predictions about empirical response times in this learning task. Firstly, response times (as well as accuracy) should be a function of both the difference in Q-values, and the inverse temperature parameter. The former is because drift rate equals the difference between the two Q-values. When choice options are similar (and the Q-value difference is small), the drift rate is low, which increases the average amount of time it takes to reach a response boundary. A negative relation between Q-value differences and response times has indeed been described in the literature (e.g., Frank et al., 2009; Krajbich et al., 2015; and applications of SSMs to value-based decisions). Furthermore, a positive relation between choice times and the inverse temperature parameter is implied because higher inverse temperatures are associated with higher threshold settings, which implies that participants accumulate *more* evidence before they commit to a decision. We are not aware of any literature specifically testing this (while simultaneously accounting for the influence of Q-value differences), but this likely holds because of the reported fits of combined RL-DDMs, which are reviewed below. Finally, the assumption that the evidence accumulation start point is unbiased towards either choice alternative entails that the response time distributions for both choice options are always symmetric ('correct' and 'wrong' answers are equally fast). There is at least some evidence to support this idea (Frank et al., 2015; Pedersen et al., 2017; but see Sewell et al., 2019). However, asymmetric response time distributions could be accounted for with a start point shift and/or the inclusion of between-trial variabilities in drift rate and start point (e.g., Ratcliff and Tuerlinckx, 2002). While no longer formally equivalent, the DDM choice function would still approximate soft-max under such a parametrisation.

3. Linking RL and SSMs: Empirical work

A rigorous test for the relation between learning dynamics and response times is to fit a combined RL-SSM[3] to empirical data (Box 1 provides a general overview of how to fit a joint RL-SSM), and to assess how well the model is able to capture all aspects of the data: The shape of response time distributions for all response types, and changes in response time and accuracy over the course of the experiment. Such tests have been provided by a series of recent studies (Fontanesi et al., 2019a, 2019b; Frank et al., 2015; Luzardo et al., 2017; Millner et al., 2018; Pedersen et al., 2017; Sewell et al., 2019), and additional studies illustrate that the approach also helps tackling problems of parameter recovery (Shahar et al., 2019; see also Ballard and McClure, 2019). Below, we review studies that formally link dynamics from a learning model with a sequential sampling model, using data from value-based decision-making tasks (e.g., instrumental learning tasks) that would traditionally have been analyzed with soft-max. However, comparable approaches have also been used in the perceptual decision-making domain (Yu and Cohen, 2009; Zhang et al., 2014), and many of the benefits advocated in the current paper apply to that domain as well.

The first study explicitly linking trial-to-trial evolution of expected values with drift rate in a DDM was provided by Frank and colleagues (2015), although they did not use a temporal difference learning rule but a Bayesian ideal observer model instead. In this simultaneous EEG-fMRI study, participants performed a probabilistic category learning task (Frank et al., 2004, Fig. 3A). By jointly modelling learning and decision-making processes, the authors showed that both the drift rate and threshold were modulated by expected value. Furthermore, in a model-based cognitive neuroscience analysis, they showed that decision thresholds were modulated by subthalamic nucleus (a small nucleus deep in the brain) activity, as well as by the dorsolateral prefrontal cortex but only under decision conflict.

Pedersen and colleagues (2017) further developed the cognitive modelling of combining learning and the DDM, and used the SARSA learning rule instead of a Bayesian ideal observer model. They showed that this combined RL-DDM was able to account for both response time distributions and learning over time in data from a probabilistic selection task. A model comparison between various RL-DDM specifications showed that a linear mapping between the Q-value difference and drift rate, combined with a time-varying threshold and different learning rates for updates after positive and negative prediction errors provided the best fit to their data. They then applied this model to data of patients diagnosed with attention deficit/hyperactivity disorder to compare their behavior on and off medication, and show that various RL-DDM parameters, including the boundary separation and learning rates, were affected by ADHD medication. Interestingly, the authors demonstrated that the model was able to fit most of the data well, but showed misfits in two aspects: It overestimated accuracy for difficult decisions; and overestimated the differences in accuracy between choice difficulty levels at the end of the experiment, while underestimating these differences in the beginning of the experiment (see Pedersen et al., 2017; their Fig. 3).

[3] Throughout the manuscript, we use RL-SSM when the sentence refers to the entire class of SSMs, and RL-DDM when it refers to an RL-SSM using the DDM specifically.

Box 1
Fitting an RL-SSM

Model fitting entails finding a set of model parameters that maximize the likelihood of the data under the model. There are many ways to optimize a set of parameters, including frequentist methods such as SIMPLEX (Nelder and Mead, 1965), differential evolution optimization (Price et al., 2006), and particle swarm optimization (Clerc, 2010), and Bayesian methods such as MCMC sampling (e.g., Ter Braak, 2006). Obtaining the likelihood of an individual subject's data under an RL-SSM depends on the update rule, the linking function, and the SSM, but in a general form can be described in the following pseudo-code:

1. Assume recursive update rule $\mathcal{f}(Q|\alpha)$, sequential sampling model $\mathcal{L}(RT, choice|\theta)$, and linking function $\mathcal{g}(Q, \theta|\psi)$
2. Generate proposal joint set of parameters α, θ, ψ
3. **for** $1 \leq n \leq N$ **do**
4. **for** $1 \leq k \leq K$ **do**
5. **if** $n = 1$ **then**
6. Initialize $Q_{k,n} = Q_{k,1} = 0.5$
7. **else**
8. Calculate $Q_{k,n}$ with update rule $f(Q_{k,n-1}|\alpha)$
9. **end if**
10. **end for**
11. Calculate θ_n with linking function $\mathcal{g}(Q_{n,\theta}|\psi)$
12. Calculate likelihood $(\mathcal{L}(RT_n, choice_n|\theta_n)$
13. **end for**
14. Calculate $\sum_{n=1}^{N} \log \mathcal{L}(RT_n, choice_n|\theta_n)$ as a measure of model fit to entire data set

Steps 2–14 are repeated using optimization algorithms until convergence. In the pseudo-code, α is a set of parameters of the update rule (e.g., the learning rate in Equation (1)), θ a set of parameters of the sequential sampling model (e.g., non-decision time, start point, and threshold of the DDM), and ψ a set of parameters of the linking function (e.g., a slope parameter under the assumption of a linear relation between drift rates and Q-value differences). The initial value of Q is set to 0.5, formalizing that the subject's expected value of each choice option is unbiased at 0.5 (with "reward" being +1 and no reward +0) at the start of the experiment. Alternatively, this initial value could be estimated as a free parameter as well. *N* is the number of trials, and *K* the number of choice options under consideration.

The overall procedure is highly comparable to fitting an RL model with soft-max. The crucial difference lies in steps 11 and 12: A linking function is required, and the likelihood function is the sequential sampling model instead of soft-max. Compared to fitting a traditional SSM, steps 4–11 are added, whereas traditional SSM fitting requires only a single set of proposal parameters θ for the entire data set.

A similar study by Fontanesi and colleagues (2019a) aimed to extend hallmark findings in value-based decision making (i.e., the effects of choice difficulty, value magnitude, and value difference) to the context of learning. Various RL, DDM, and integrated RL-DDMs were fit to the data, after which both qualitative and quantitative model comparisons showed that only the integrated RL-DDM was able to account for all aspects of the data (see Fig. 3B). In order to overcome the issue of overestimating choice accuracy for difficult choices reported by Pedersen et al. (2017), a non-linear link function between expected values and drift rate was required. Furthermore, the effect of value magnitude was accounted for by allowing the threshold effect to vary as a function of the mean expected value (c.f. Frank et al., 2015; Ratcliff and Frank, 2012), indicating that participants responded less cautiously when a larger gain could be earned.

In another study, Fontanesi and colleagues (2019b) applied the RL-DDM framework to understand how context effects influence response times and accuracy in an instrumental learning task. They re-analyzed data from four experiments in which context was manipulated by altering the valence of feedback: Participants had to learn to maximize gains for some choice options (reward learning), and minimize losses for others (avoidance learning). Furthermore, participants received either feedback for the rewards associated with both choice options (c.f. Fig. 2A), or only the choice option that was actually chosen. Using the RELATIVE model (Palminteri et al., 2015) as a learning model, they showed that the valence of feedback consistently affected non-decision time: In trials in which participants had to make a decision in order to gain a reward, non-decision time was lower than in trials in which participants had to decide in order to avoid a loss. Furthermore, giving full feedback instead of partial feedback increased the drift rate and the threshold, and decreased the non-decision times of choices. Finally, in line with earlier findings (Frank et al., 2015), decision conflict (the inverse of difficulty) affected the threshold. These results are especially interesting since they show that the decision-making process is not only affected by the expected reward, but also by the state that the decision-maker is in.

Sewell and colleagues (2019) took a related approach to modelling choice data from a probabilistic learning task, using an associative network (ALCOVE; Kruschke, 1992) instead of a classical reinforcement learning model to model learned changes in associative strength between stimuli and outcomes. This model was able to account for changes in choice probabilities and response latencies due to learning, as well as the effect of choice difficulty. To do so, a non-linear linking function between expected values and drift rates was required. However, they note that the model showed minor misfits in two aspects of the data: It could not capture the observed asymmetry between response latencies associated with correct and incorrect responses (errors were consistently slower; see Fig. 3C), and the model predicted a higher skewness in the response time distributions than observed.

An interesting application of an integrated RL-DDM was given by Millner and colleagues (2018), who used this framework to improve our understanding of how aversive Pavlovian biases shape choice behavior. Such biases are thought to be the result of a hard-wired response to avoid aversive stimuli, and have been previously associated with response inhibition. Using their Go-Nogo paradigm combined with RL-DDM modelling, the authors showed not only that Pavlovian biases can also promote active behavior (rather than only response inhibition), but that the bias was best understood as a *response* bias (or start point bias) in decision-making. That is, when faced with an aversive stimulus, a Pavlovian bias reduces the amount of evidence participants require to decide towards the response that minimizes the influence of aversive stimuli, regardless of whether this response is active or passive.

It is important to highlight that in the studies mentioned above, the

results could not have been obtained without jointly modelling the learning *and* decision parts of the data. Furthermore, while all models appear to be able to capture fundamental aspects of the decision and learning processes, misfits remain, providing a challenge and opportunity for future model development.

4. Benefits of integration

The previous section reviewed the current work on combining reinforcement learning models with SSMs, providing several illustrations of how such integrated models can improve our understanding of learning and decision making simultaneously. Here, we generalize beyond individual use-cases, and argue that integration of the two modelling traditions is promising for both cognitive modelling and model-based cognitive neuroscience.

4.1. Methodological benefits: Improved parameter recovery

Before parameters can be interpreted or related to neural measures, it is crucial that the true data-generating parameter values can be recovered from the model when it is fit to data (Moran, 2016; Spektor and Kellen, 2018). Good parameter recovery involves reliably and accurately estimating the true parameter values that generated a set of observed or simulated data, and provides a minimally necessary condition that must be met before one can make inferences about latent cognitive processes from model parameters. Although it may seem trivial, good recovery is a known issue for both RL models (Spektor and Kellen, 2018; Wetzels et al., 2010) and SSMs (Boehm et al., 2018; Miletić et al., 2017; Ratcliff and Tuerlinckx, 2002; Van Ravenzwaaij and Oberauer, 2009).

It is promising to see that current RL-SSM studies (Fontanesi et al., 2019a; Pedersen et al., 2017) recognized the importance of good parameter recovery, and explicitly studied the recovery properties of the models employed in their studies. These RL-DDMs recovered remarkably well given the numbers of trials included. Another recent study showed that combining RL models with the DDM significantly improves recovery compared to more common (relatively complex) RL mixture models (Shahar et al., 2019; see also Ballard and McClure, 2019). The reason behind these good recovery properties is that including response times increases the information content of the data, and the amount of constraint offered by the data. In fact, in modelling the two-stage decision task (Daw et al., 2011) using a common (choice-only) RL mixture model of model-based and model-free learning, Shahar et al. (2019) reported that very high numbers of trials (>1000) were required to obtain good parameter recovery properties for the crucial mixture proportion parameter, against only 200 trials for a combined RL-DDM. In this light, modelling the data using a combined RL-SSM may even be considered *necessary* for such relatively complex RL models. Good recovery with small numbers of trials is especially beneficial for model-based cognitive neuroscience, where collection of large numbers of trials is too time consuming and/or costly.

It would be interesting to see whether including learning in modelling also improves recovery of commonly-used SSM parameters. For example, parameters designed to model between-trial variabilities in the DDM (Boehm et al., 2018), urgency signals (Evans et al., 2019; Voskuilen et al., 2016), and evidence leakage and inhibition effects (Miletić et al., 2017) often show poor recovery due to trade-offs with other parameters. The additional trial-by-trial constraints on evidence accumulation rates likely decreases covariances between parameters, which could aid parameter recovery.

4.2. Benefits for reinforcement learning models

4.2.1. Explaining response times

From the RL perspective, a process-agnostic choice rule is replaced with a psychologically-principled process model that describes how decisions arise from value representations that change over time due to learning. The formalization of decision making as an accumulate-to-threshold process allows for the explanation of a whole new dimension of the behavioral data in terms of explaining the shape of entire response time distributions.

One may wonder why modelling response time latencies is important when choices (and changes thereof) are of primary interest to a researcher. The reason is that response latencies and choice behavior have long been known to trade-off (for early work, e.g., Wickelgren, 1977), and this trade-off is under voluntary control of the participant (for reviews, Bogacz et al., 2010; Heitz, 2014). As a consequence, studying changes in choice behavior without considering the concurrent changes in response latencies could lead to biased conclusions. For example, an increase in accuracy over time could indicate that a

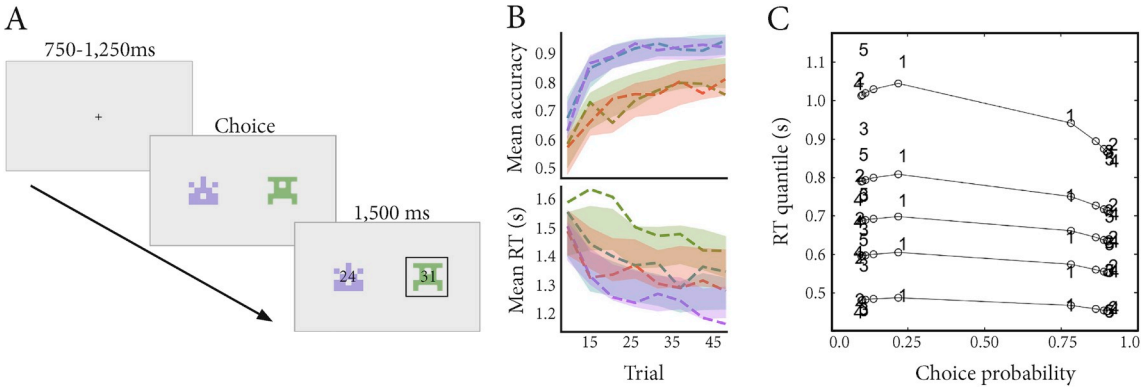

Fig. 3. A) Typical instrumental learning task, in which a participant has to decide between two stimuli, each associated with a probabilistic reward. By trial and error, the participant needs to learn which stimuli lead to greatest reward. Reproduced from Fontanesi et al. (2019a, 2019b) with permission. B) Change in accuracy (top panel) and mean RT (bottom panel) over time. Dashed lines are data, shaded area corresponds to 95% Bayesian credible interval of the RL-DDM predictions, and colors indicate choice difficulty. The RL-DDM was able to capture the increase in accuracy and decrease in mean RT over time. Figure adapted from Fontanesi et al. (2019a, 2019b) with permission. C) Quantile probability plot of fit of entire response time distributions for five learning blocks in Sewell et al. (2019), only for choices where the probability of reward for both choices options were 0.8 and 0.2. In this plot (see Ratcliff and Smith, 2011 for a more detailed explanation of this type of graph), circles indicate the model predictions, and numbers correspond to data for five learning blocks. The x-axis of the numbers and circles indicates the proportion of choices (i.e., right side circles/numbers are correct answers, left side are incorrect answers), and the y-axis indicates the 10th, 30th, 50th, 70th, and 90th percentile of the response times. The combined learning-DDM was able to capture the overall response time distributions, although some misfits are present (e.g., an underestimation of the incorrect response times in learning blocks 1, 3, and 5). Reproduced with permission from Springer: Psychonomic Bulletin & Review from Sewell et al. (2019). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

participant improves at a task (i.e., a practice effect), but this conclusion only holds if response times decrease (or remain the same) as learning progresses. Increases in response times would suggest that participants increase their response caution instead. Conversely, there may be no change in choice behavior over time (which may lead a researcher to believe that no learning took place), but a striking decrease in response times. This could indicate that the participant *did* learn and simultaneously decreased response caution, leading to similar accuracy over the entire course of learning. Combined RL-SSMs prevent such inferential biases by considering *both* choices and response latencies.

4.2.2. Interpreting the soft-max inverse temperature parameter

RL-SSM can offer alternative interpretation of the soft-max inverse temperature parameter. This parameter is typically interpreted in terms of the exploitation/exploration trade-off (e.g., Daw et al., 2006), with higher values indicating that participants exploit current knowledge of the likely rewards associated with each response alternative (i.e., participants stick with responses known to result in rewards), and lower values indicating that participants explore the environment more (i.e., participants test out alternative responses to see whether they result in rewards). Alternatively, the inverse temperature parameter has been interpreted in terms of sensitivity to reward (e.g., Fontanesi et al., 2019a; Pedersen et al., 2017), since higher values indicate that the currently expected reward has a more deterministic influence on choice behavior. The RL-DDM specifically, in contrast, offers a different interpretation of this parameter (and its behavioral effects) in terms of the speed-accuracy trade-off (SAT) (Tuerlinckx and De Boeck, 2005): Some participants are slower and more accurate than others because they inform their decisions based on more evidence (i.e., they set higher response thresholds). Unlike the traditional interpretations of the inverse temperature parameter, this alternative interpretation thus makes formal predictions about response times. These predictions furthermore entail more than a mere increase or decrease in the mean response times, but are about the shapes of the entire response time distributions of both correct and incorrect responses. Future experiments, for example using a classical speed/accuracy trade-off manipulation, should test whether empirical data supports this interpretation.

On top of explicitly modelling SAT settings, RL studies may further benefit from the richness of latent variables estimated using SSMs, such as the amount of time it takes for participants to perform early perceptual processes and initiate motor responses (i.e., non-decision time), potential response biases, and the efficiency with which participants process information (i.e., accumulation rates). One field where this may be of special interest is computational psychiatry research (Wiecki et al., 2015), since SSMs allow for a more detailed description of differences in decision-making processes between healthy and clinical populations, and the effect of medication. Pedersen et al. (2017) already provided an example of how a combined RL-DDM was used to analyze the complex effects of medication on response thresholds, non-decision times and learning rates in patients with ADHD. These complex effects could not be captured by the inverse temperature parameter of soft-max.

4.2.3. Model selection

Model selection entails finding the most parsimonious trade-off between quality of fit and model complexity. However, in some cases, different models of the same complexity can make the exact same predictions for choice data, which makes it difficult if not impossible to distinguish between these models. In such a case, response times can provide additional and sometimes crucial constraint to inform model selection. An example is the study of context effects on preferential decision-making, where it was shown that some models can only be distinguished on the basis of the combination of choice and response time data (Busemeyer et al., 2019; Molloy et al., 2019; Turner et al., 2018b).

Mixture models form another example where response times can be especially informative (Van Maanen et al., 2016, 2014). Mixture models in decision-making propose that observed data are the result of not a single decision-making process, but of a mixture between two or more processes (or cognitive strategies). A well-known example in reinforcement learning are the model-free and model-based control systems (Daw and Dayan, 2014). Model-free control assumes that Q-values are computed solely by trial-and-error (i.e., Equation (1)), while model-based control assumes that Q-values are explicitly computed based on an internal cognitive model of the world. Observed choice behavior is thought to be a mixture of both types of control, and the mixture parameter, that denotes how reliant a subject is on either of these strategies, is often analyzed (Daw et al., 2011). However, as also discussed above, the amount of information in choice data alone is often too limited to reliably estimate parameters of such a mixture model (Shahar et al., 2019).

The SSM framework offers exciting new ways to model competing choice strategies. One option is to assume that Q-values are a mixture of the two control systems, and link the difference in Q-values to drift rates of an SSM (Shahar et al., 2019). Another option would be to model the different strategies as a race between accumulators, as is typically done in race models such as the linear, ballistic accumulator (LBA) model (Brown and Heathcote, 2008). This way, model selection between RL-SSMs provides a way to test not only for the existence of multiple choice strategies in learning tasks, but also how exactly these strategies interact and/or compete.

4.3. Benefits for sequential sampling models

4.3.1. An explicit theory of drift rates

From the perspective of SSMs, a major benefit from the integration is that an explicit theory of drift rates is incorporated into the model, specifying exactly what constitutes the evidence that is being accumulated. By equating (or non-linearly linking) drift rates with Q-value differences, RL-SSMs propose that evidence accumulation is driven by (differences in) expected value. In simple instrumental learning tasks, where model-free control drives choice behavior, Q-values are thought to be stored in procedural memory (Gershman and Daw, 2017). By modelling such a task using a RL-SSM, the time it takes to perform additional cognitive processes such as stimulus identification will be accounted for by the non-decision time parameter, and variability in response latencies are a consequence of the noise in Q-value accumulation. In more complex tasks (e.g., the two-stage decision task, Daw et al., 2011), Q-values can either be thought of as a mixture of multiple control systems, or the decision-making process can be modelled as a race between such control systems, as proposed above.

4.3.2. Single trial parameter estimates

RL models offer trial-by-trial estimates of expected value and prediction errors, which have been used in model-based cognitive neuroscience to make inferences about trial-by-trial variability in neural measures. This approach involves using expected values or prediction errors as parametric modulators in analyses of neural data, and has proven to be very powerful in identifying the brain areas putatively involved in computing and representing these variables (O'Doherty et al., 2007).

The default SSM framework, by contrast, does not provide a single parameter estimate per trial. Instead, one set of parameters per condition is typically estimated, and it is assumed that all responses within this condition are independent draws from a single distribution. Apart from the fact that this assumption is difficult to entertain (as evidenced, for example, by the phenomenon of post-error slowing; Dutilh et al., 2012), it also limits the methodology of model-based cognitive neuroscience. Since only one parameter estimate is available per condition and subject, assuming that all choice processes within that condition are identical (except for within-trial noise), there is no explicit model of trial-by-trial dynamics that can be used to analyze within-subject dynamics in neural measures. As an alternative approach, between-subject

dynamics are often analyzed, showing that the size of a between-condition BOLD-response contrast in certain brain areas (e.g., anterior striatum) covaries across subjects with corresponding changes in model parameters (e.g., threshold) (Forstmann et al., 2010, 2008; Mulder et al., 2012). While insightful, between-subject correlations between parameter estimates and neural data offer substantially less detailed descriptions of behavior than within-subject analyses, and effects that may be present at the group level (i.e., based on aggregated data) do not necessarily translate into commensurate effects at the individual level (e.g., Simpson's paradox; Kievit et al., 2013). To overcome this limitation, substantial effort has been put into developing methods to obtain trial-by-trial parameter estimates, either using behavioral data alone (Gluth and Meiran, 2019; Van Maanen et al., 2011), or by making use of behavioral and neural data together, as in joint modelling approaches (Turner et al., 2019b, 2019a; 2018a, 2017b, 2015).

By explicitly linking drift rates to expected values, unified RL-SSM models provide trial-by-trial estimates of drift rates. Furthermore, the additional constraint offered by the data has been shown to allow for a more fine-grained description of the decision-making behavior, for example by including trial-by-trial estimates of threshold (Fontanesi et al., 2019a, 2019b; Frank et al., 2015; Pedersen et al., 2017) as well. Additional trial-order effects such as post-error slowing (Dutilh et al., 2012) could potentially be modelled by allowing the threshold to vary as a function of prediction errors. Together, the combination of RL and SSM models increases the level of detail in the cognitive model, allowing for more powerful inferences at the single-trial level.

4.4. Benefits for model-based cognitive neuroscience

An integrated RL-SSM is likely closer to the data-generating model of human behavior than either model alone. This is crucial for model-based cognitive neuroscience, because by placing the cognitive model "in the middle" of the behavioral and neural data, the analyses of neural data are not only driven by the latent cognitive processes a model is able to identify, but also constrained by a model's limitations: Potential model misspecifications or biases propagate into the model-based analyses, limiting or biasing the final conclusions. In order to draw valid conclusions about the brain and cognition, the cognitive models employed should be as unbiased as possible with respect to the data-generating process, within the limits of what the data allow for. It is a matter of model development (and competitive model selection) to discover which model specifications are closest to the ground truth.

5. Future directions

In model-based cognitive neuroscience, the quality of inferences in neural functioning depends on how well cognitive models approximate the actual underlying cognitive processes, rendering cognitive model development of crucial importance to the field. In this light, the recent efforts in cognitive modelling toward integrating reinforcement learning and sequential sampling traditions into a unified theoretical framework are highly exciting.

The integration of the RL and SSM frameworks poses new avenues for future research. Firstly, what is the best way to model choice behavior in learning tasks? The currently reviewed studies used the DDM as a choice function, but despite its ability to capture the main aspects of the data, various misfits were also reported. Since race models such as the linear ballistic accumulator model (LBA; Brown and Heathcote, 2008) have been very successful in explaining choice behavior in both perceptual and value-based decisions (Busemeyer et al., 2019; Rodriguez et al., 2015, 2014; Turner et al., 2018b), it would be informative to see if they can help overcome the misfits reported earlier. Furthermore, unlike the DDM (but see Krajbich and Rangel, 2011), race models can easily be extended to multi-alternative choice tasks. However, there are various other SSMs such as the leaky competing accumulator model (Usher and McClelland, 2001), urgency models (Hawkins et al., 2015; Thura and Cisek, 2016), and racing diffusion models (e.g., Boucher et al., 2007; for a more complete overview, see Bogacz et al., 2006, their Fig. 2; and Ratcliff et al., 2016, their Fig. 2). It is crucial to test which SSMs provide the best fit to decision-making data in learning contexts.

Cognitive model development also entails testing to what extent learning and decision-making processes operate independently, and to what extent they interact. The current literature consistently finds that drift rates fluctuate as a function of reward expectations, although it remains a topic of discussion whether this mapping is linear or nonlinear. Furthermore, some studies (Fontanesi et al., 2019a; Frank et al., 2015) find an effect of expected value on decision thresholds. On top of this, it may be that the start point of evidence accumulation can be biased due to learning effects (e.g., Millner et al., 2018), which could potentially also explain the slowness of errors relative to correct responses reported by Sewell et al. (2019). A related interesting question is how SAT settings influence learning. Sewell et al. (2019) suggest that the learning rate could be affected by imposing speed stress in the decision-making phase. The underlying reasoning is that speed stress generally causes participants to inform their decisions based on less evidence (i.e., low thresholds) (Heitz, 2014), which increases the likelihood of an error and thereby decreases the information content in an error compared to errors that occur when choices are based on a lot of evidence (i.e., high thresholds).

Similar to progressing our understanding of how the cognitive systems of learning and decision making interact, a model-based cognitive neuroscience using RL-SSMs can also improve our understanding of how the neural systems underlying learning and decision-making interact. Both in the RL framework (O'Doherty et al., 2017) as in the SSM framework (Mulder et al., 2014), much progress has been made in our understanding of the neural systems underlying learning and perceptual decision making separately. These brain networks at least partly overlap. One example structure is the striatum, of which different parts have been implicated in processing reward prediction errors (e.g., Haruno and Kawato, 2006; O'Doherty et al., 2003) and learning of action-outcome associations (Balleine et al., 2007; Kim et al., 2009), as well as response caution settings (Forstmann et al., 2008; Van Maanen et al., 2011). A second example is the parietal cortex, which has been implicated in perceptual evidence accumulation (e.g., Shadlen and Newsome, 2001), as well as value-based decision-making (e.g., Platt and Glimcher, 1999) and state prediction error updating (Gläscher et al., 2010). Using methods similar to Frank and colleagues (2015), or potentially joint modelling (Turner et al., 2019a, 2019b), it would be interesting to test whether the similarity in neural systems alludes to the same cognitive processes playing a role in both RL and decision-making, or that different cognitive processes are underpinned by the same neural substrates.

The suggestions above only scratch the surface of the plethora of modelling options in combining RL and SSMs. This combinatorial explosion of modelling options simultaneously forms a challenge for future research. From the RL framework, multiple learning rules can be taken, as well as additional mechanisms (mentioned briefly above) such as multiple learning rates, eligibility traces, choice perseveration, and mixture modelling. Further, as detailed above, there exists a multitude of SSMs, including the DDM with various parametrizations, and race models such as racing diffusion accumulators and the LBA, and urgency models. Finally, as already shown by the current literature, there are multiple linear and nonlinear linking functions possible between RL model dynamics and SSM parameters. Again, this landscape of options is exciting, but risks post-hoc modelling decisions and overfitting issues. Selecting the most theoretically informative combination of RL and SSM mechanisms and the most appropriate linking functions from this broad landscape will be an important challenge for future model development and selection.

A second (and related) challenge is to make quantitative comparisons between models that vary in the dimensionality of the data they

explain. More specifically, traditional RL models using soft-max treat data as univariate (i.e., choices only), whereas combined RL-SSMs treat data as multivariate (i.e., choices with associated response latencies). An unpractical consequence of this difference in dimensionality is that likelihood estimates of soft-max and SSMs (and, by extension, model comparison metrics such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), and their Bayesian extensions such as the DIC (Spiegelhalter et al., 2002)) are informed by different data, and cannot be directly compared. A potential option to overcome this issue is to resort to cross-validation techniques (e.g., Ahn et al., 2008; Steingroever et al., 2014). In such techniques, the models are fit to part of the data, and subsequently used to generate predictions on choice behavior in a separate part of the data. With RL-SSMs, it would be possible to repeatedly fit the model to the first n trials, and predict the choice and response time for trial n+1. The accuracy of such predictions can be compared and interpreted as a measure of generalizability. Apart from these relative model selection criteria, it is important to also assess the absolute quality of fit (Palminteri et al., 2017), to ensure that the model is able to capture key phenomena in the data.

Neither of these challenges poses a fundamental hurdle to integrating the cognitive model classes. Moreover, although the articles reviewed here focus on value-based decision-making, very similar benefits also apply to the perceptual learning domain (e.g., Diaz et al., 2017; Lak et al., 2017). Careful model development, combined with experimental work and tests of specific influence, has the potential to offer the field of model-based cognitive neuroscience a powerful new tool for measuring and interpreting behavioral and neural data within a single theoretical framework encompassing a variety of cognitive processes.

Acknowledgements

This work was supported by a grant from the Netherlands Organisation for Scientific Research (NWO; grant number 016.Vici.185.052; BUF).

References

Ahn, W.Y., Busemeyer, J.R., Wagenmakers, E.J., Stout, J.C., 2008. Comparison of decision learning models using the generalization criterion method. Cogn. Sci. 32, 1376–1402. https://doi.org/10.1080/03640210802352992.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caski, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Ballard, I.C., McClure, S.M., 2019. Joint modeling of reaction times and choice improves parameter identifiability in reinforcement learning models. J. Neurosci. Methods 317, 37–44. https://doi.org/10.1016/j.jneumeth.2019.01.006.

Balleine, B.W., Delgado, M.R., Hikosaka, O., 2007. The role of the dorsal striatum in reward and decision-making. J. Neurosci. 27, 8161–8165. https://doi.org/10.1523/jneurosci.1554-07.2007.

Boehm, U., Annis, J., Frank, M.J., Hawkins, G.E., Heathcote, A., Kellen, D., Krypotos, A. M., Lerche, V., Logan, G.D., Palmeri, T.J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J.J., Voss, A., Wiecki, T.V., Matzke, D., Wagenmakers, E.J., 2018. Estimating across-trial variability parameters of the diffusion decision model: expert advice and recommendations. J. Math. Psychol. 87, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004.

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. Psychol. Rev. 113, 700–765. https://doi.org/10.1037/0033-295X.113.4.700.

Bogacz, R., McClure, S.M., Li, J., Cohen, J.D., Montague, P.R., 2007. Short-term memory traces for action bias in human reinforcement learning. Brain Res. 1153, 111–121. https://doi.org/10.1016/j.brainres.2007.03.057.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B.U., Nieuwenhuis, S., 2010. The neural basis of the speed-accuracy tradeoff. Trends Neurosci. 33, 10–16. https://doi.org/10.1016/j.tins.2009.09.002.

Boucher, L., Palmeri, T.J., Logan, G.D., Schall, J.D., 2007. Inhibitory control in mind and brain:an interactive race model of countermanding. Saccades 114, 376–397. https://doi.org/10.1037/0033-295X.114.2.376.

Brown, S.D., Heathcote, A., 2008. The simplest complete model of choice response time: linear ballistic accumulation. Cogn. Psychol. 57, 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002.

Busemeyer, J.R., Gluth, S., Rieskamp, J., Turner, B.M., 2019. Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. Trends Cogn. Sci. 23, 251–263. https://doi.org/10.1016/j.tics.2018.12.003.

Christakou, A., Gershman, S.J., Niv, Y., Simmons, A., Brammer, M., Rubia, K., 2013. Neural and psychological maturation of decision-making in adolescence and young adulthood. J. Cogn. Neurosci. 25, 1807–1823. https://doi.org/10.1162/jocn_a_00447.

Clerc, M., 2010. Particle Swarm Optimization, vol. 93. John Wiley & Sons, Inc.

Daw, N.D., Dayan, P., 2014. The algorithmic anatomy of model-based evaluation. Philos. Trans. R. Soc. Biol. Sci. 369 https://doi.org/10.1098/rstb.2013.0478.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027.

Daw, N.D., Kakade, S., Dayan, P., 2002. Opponent interactions between serotonin and dopamine. Neural Netw. 15, 603–616. https://doi.org/10.1016/S0893-6080(02)00052-7.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441, 876–879. https://doi.org/10.1038/nature04766.

Dezfouli, A., Balleine, B.W., 2013. Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput. Biol. 9 https://doi.org/10.1371/journal.pcbi.1003364.

Diaz, J.A., Queirazza, F., Philiastides, M.G., 2017. Perceptual learning alters post-sensory processing in human decision-making. Nat. Hum. Behav. 1 https://doi.org/10.1038/s41562-016-0035.

Doll, B.B., Duncan, K.D., Simon, D.A., Shohamy, D., Daw, N.D., 2015. Model-based choices involve prospective neural activity. Nat. Neurosci. 18, 767–772. https://doi.org/10.1038/nn.3981.

Dutilh, G., Vandekerckhove, J., Forstmann, B.U., Keuleers, E., Brysbaert, M., Wagenmakers, E.J., 2012. Testing theories of post-error slowing. Atten. Percept. Psychophys. 74, 454–465. https://doi.org/10.3758/s13414-011-0243-2.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., Wagenmakers, E.J., 2009. A diffusion model decomposition of the practice effect. Psychon. Bull. Rev. 16, 1026–1036. https://doi.org/10.3758/16.6.1026.

Evans, N.J., Brown, S.D., Mewhort, D.J.K., Heathcote, A., 2018. Refining the law of practice. Psychol. Rev. 125, 592–605. https://doi.org/10.1037/rev0000105.

Evans, N.J., Trueblood, J.S., Holmes, W.R., 2019. A parameter recovery assessment of time-variant models of decision-making. Behav. Res. Methods. https://doi.org/10.3758/s13428-019-01218-0.

Fontanesi, L., Gluth, S., Spektor, M.S., Rieskamp, J., 2019. A reinforcement learning diffusion decision model for value-based decisions. Psychon. Bull. Rev. https://doi.org/10.3758/s13423-018-1554-2.

Fontanesi, L., Palminteri, S., Lebreton, M., 2019. Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. Cognit. Affect Behav. Neurosci. 19, 490–502. https://doi.org/10.3758/s13415-019-00723-1.

Forstmann, B.U., Anwander, A., Schäfer, A., Neumann, J., Brown, S.D., Wagenmakers, E.-J., Bogacz, R., Turner, R., 2010. Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. Proc. Natl. Acad. Sci. U.S.A. 107, 15916–15920. https://doi.org/10.1073/pnas.1004932107.

Forstmann, B.U., Dutilh, G., Brown, S.D., Neumann, J., von Cramon, D.Y., Ridderinkhof, K.R., Wagenmakers, E.-J., 2008. Striatum and pre-SMA facilitate decision-making under time pressure. Proc. Natl. Acad. Sci. U.S.A. 105, 17538–17542. https://doi.org/10.1073/pnas.0805903105.

Forstmann, B.U., Ratcliff, R., Wagenmakers, E.-J., 2016. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. Annu. Rev. Psychol. 67, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645.

Forstmann, B.U., Wagenmakers, E.-J., 2015. An Introduction to Model-Based Cognitive Neuroscience. Springer. https://doi.org/10.1007/978-1-4939-2236-2.

Forstmann, B.U., Wagenmakers, E., Eichele, T., Brown, S.D., Serences, J.T., 2011. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract ? Trends Cogn. Sci. 15, 272–279. https://doi.org/10.1016/j.tics.2011.04.002.

Frank, M.J., Doll, B.B., Oas-Terpstra, J., Moreno, F., 2009. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci. 12, 1062–1068. https://doi.org/10.1038/nn.2342.

Frank, M.J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T.V., Cavanagh, J.F., Badre, D., 2015. fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. J. Neurosci. 35, 485–494. https://doi.org/10.1523/JNEUROSCI.2036-14.2015.

Frank, M.J., Seeberger, L.C., O'Reilly, R.C., 2004. By carrot or by stick:cognitive reinforcement learning in parkinsonism author (s) :Michael J . Frank , Lauren C . Seeberger and Randall C. O' Reilly. Science 306 (80), 1940–1943.

Gershman, S.J., 2015. Do learning rates adapt to the distribution of rewards? Psychon. Bull. Rev. 22, 1320–1327. https://doi.org/10.3758/s13423-014-0790-3.

Gershman, S.J., Daw, N.D., 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu. Rev. Psychol. 68, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625.

Gläscher, J.P., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595. https://doi.org/10.1016/j.neuron.2010.04.016.

Gläscher, J.P., O'Doherty, J.P., 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscipl. Rev. Cogn. Sci. 1, 501–510. https://doi.org/10.1002/wcs.57.

Gluth, S., Meiran, N., 2019. Leave-one-trial-out, LOTO, a general approach to link single-trial parameters of cognitive models to neural data. Elife 8, 1–39. https://doi.org/10.7554/eLife.42607.

Gold, J.I., Shadlen, M.N., 2001. Neural computations that underlie decisions about sensory stimuli. Trends Cogn. Sci. 5, 10–16. https://doi.org/10.1016/S1364-6613(00)01567-9.
Haruno, M., Kawato, M., 2006. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. J. Neurophysiol. 95, 948–959. https://doi.org/10.1152/jn.00382.2005.
Haughey, H.M., Hutchison, K.E., Curran, T., Frank, M.J., Moustafa, A.A., 2007. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. 104, 16311–16316. https://doi.org/10.1073/pnas.0706111104.
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. J. Neurosci. 35, 2476–2484. https://doi.org/10.1523/JNEUROSCI.2410-14.2015.
Heitz, R.P., 2014. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. Front. Neurosci. 8, 1–19. https://doi.org/10.3389/fnins.2014.00150.
Kievit, R.A., Frankenhuis, W.E., Waldorp, L.J., Borsboom, D., 2013. Simpson's paradox in psychological science: a practical guide. Front. Psychol. 4, 1–14. https://doi.org/10.3389/fpsyg.2013.00513.
Kim, H., Sul, J.H., Huh, N., Lee, D., Jung, M.W., 2009. Role of striatum in updating values of chosen actions. J. Neurosci. 29, 14701–14712. https://doi.org/10.1523/jneurosci.2728-09.2009.
Kool, W., Cushman, F.A., Gershman, S.J., 2018. Competition and cooperation between multiple reinforcement learning systems. In: Morris, R.W., Bornstein, A. (Eds.), Goal-Directed Decision Making. Elsevier, pp. 153–178. https://doi.org/10.1016/B978-0-12-812098-9.00007-3.
Krajbich, I., Armel, C., Rangel, A., 2010. Visual fixations and the computation and comparison of value in simple choice. Nat. Neurosci. 13, 1292–1298. https://doi.org/10.1038/nn.2635.
Krajbich, I., Bartling, B., Hare, T., Fehr, E., 2015. Rethinking fast and slow based on a critique of reaction-time reverse inference. Nat. Commun. 6, 1–9. https://doi.org/10.1038/ncomms8455.
Krajbich, I., Rangel, A., 2011. Multialternative Drift-Diffusion Model Predicts the Relationship between Visual Fi Xations and Choice in Value-Based Decisions, vol. 108. https://doi.org/10.1073/pnas.1101328108.
Kruschke, J.K., 1992. ALCOVE: an exemplar-based connectionist model of category learning. Psychol. Rev. 99, 22–44. https://doi.org/10.1037/0033-295X.99.1.22.
Lak, A., Nomoto, K., Keramati, M., Sakagami, M., Kepecs, A., 2017. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Curr. Biol. 27, 821–832. https://doi.org/10.1016/j.cub.2017.02.026.
Leong, Y.C., Hughes, B.L., Wang, Y., Zaki, J., 2019. Neurocomputational mechanisms underlying motivated seeing. Nat. Hum. Behav. https://doi.org/10.1038/s41562-019-0637-z.
Liu, C.C., Watanabe, T., 2012. Accounting for speed-accuracy tradeoff in perceptual learning. Vis. Res. 61, 107–114. https://doi.org/10.1016/j.visres.2011.09.007.
Luzardo, A., Alonso, E., Mondragón, E., 2017. A Rescorla-Wagner drift-diffusion model of conditioning and timing. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1005796.
Miletić, S., Turner, B.M., Forstmann, B.U., Van Maanen, L., 2017. Parameter recovery for the leaky competing accumulator model. J. Math. Psychol. 76, 25–50. https://doi.org/10.1016/j.jmp.2016.12.001.
Millner, A.J., Gershman, S.J., Nock, M.K., den Ouden, H.E.M., 2018. Pavlovian control of escape and avoidance. J. Cogn. Neurosci. 30, 1379–1390. https://doi.org/10.1162/jocn_a_01224.
Milosavljevic, M., Malmaud, J., Huth, A., 2010. The Drift Diffusion Model Can Account for the Accuracy and Reaction Time of Value-Based Choices under High and Low Time Pressure. October, vol. 5, pp. 437–449. https://doi.org/10.2139/ssrn.1901533.
Molloy, F.M., Galdo, M., Bahg, G., Liu, Q., Turner, B.M., 2019. What's in a response time?: on the importance of response time measures in constraining models of context effects. Decision 6, 171–200. https://doi.org/10.1037/dec0000097.
Moran, R., 2016. Thou shalt identify! the identifiability of two high-threshold models in confidence-rating recognition (and super-recognition) paradigms. J. Math. Psychol. 73, 1–11. https://doi.org/10.1016/j.jmp.2016.03.002.
Mulder, M.J., Van Maanen, L., Forstmann, B.U., 2014. Perceptual decision neurosciences - a model-based review. Neuroscience 277, 872–884. https://doi.org/10.1016/j.neuroscience.2014.07.031.
Mulder, M.J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., Forstmann, B.U., 2012. Bias in the brain: a diffusion model analysis of prior probability and potential payoff. J. Neurosci. 32, 2335–2343. https://doi.org/10.1523/JNEUROSCI.4156-11.2012.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313. https://doi.org/10.1093/comjnl/7.4.308.
Niv, Y., Edlund, J.A., Dayan, P., O'Doherty, J.P., 2012. Neural prediction errors Reveal a Risk-sensitive reinforcement-learning process in the human brain. J. Neurosci. 32, 551–562. https://doi.org/10.1523/jneurosci.5498-10.2012.
Nunez, M.D., Srinivasan, R., Vandekerckhove, J., 2015. Individual differences in attention influence perceptual decision making. Front. Psychol. 8, 1–13. https://doi.org/10.3389/fpsyg.2015.00018.
Nunez, M.D., Vandekerckhove, J., Srinivasan, R., 2017. How attention influences perceptual decision making: single-trial EEG correlates of drift-diffusion model parameters. J. Math. Psychol. 76B https://doi.org/10.1016/j.jmp.2016.03.003.
O'Doherty, J.P., 2014. The problem with value. Neurosci. Biobehav. Rev. 43, 259–268. https://doi.org/10.1016/j.neubiorev.2014.03.027.
O'Doherty, J.P., Cockburn, J., Pauli, W.M., 2017. Learning, reward, and decision making. Annu. Rev. Psychol. 68, 73–100. https://doi.org/10.1146/annurev-psych-010416-044216.
O'Doherty, J.P., Dayan, P., Friston, K.J., Critchley, H., Dolan, R.J., 2003. Temporal difference models and reward-related learning in the human brain. Neuron 28, 329–337.
O'Doherty, J.P., Hampton, A., Kim, H., 2007. Model-based fMRI and its application to reward learning and decision making. Ann. N. Y. Acad. Sci. 1104, 35–53. https://doi.org/10.1196/annals.1390.022.
Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual modulation of value signals in reward and punishment learning. Nat. Commun. 6 https://doi.org/10.1038/ncomms9096.
Palminteri, S., Wyart, V., Koechlin, E., 2017. The importance of falsification in computational cognitive modeling. Trends Cogn. Sci. 21, 425–433. https://doi.org/10.1016/j.tics.2017.03.011.
Pedersen, M.L., Frank, M.J., Biele, G., 2017. The drift diffusion model as the choice rule in reinforcement learning. Psychon. Bull. Rev. 24, 1234–1251. https://doi.org/10.3758/s13423-016-1199-y.
Petrov, A.A., van Horn, N.M., Ratcliff, R., 2011. Dissociable perceptual-learning mechanisms revealed by diffusion-model analysis. Psychon. Bull. Rev. 18, 490–497. https://doi.org/10.3758/s13423-011-0079-8.
Platt, M.L., Glimcher, P.W., 1999. Neural correlates of decision variables in parietal cortex. Nature 400, 233–238. https://doi.org/10.1038/22268.
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, 709–720. https://doi.org/10.1016/j.neuron.2014.03.014.
Price, K.V., Storn, R.M., Lampinen, J.A., 2006. Differential Evolution - A Practical Approach to Global Optimization, Natural Computing. Springer-Verlag.
Purcell, B.A., Heitz, R.P., Cohen, J.Y., Schall, J.D., Logan, G.D., Palmeri, T.J., 2010. Neurally constrained modeling of perceptual decision making. Psychol. Rev. 117, 1113–1143. https://doi.org/10.1037/a0020311.
Ratcliff, R., 1978. A theory of memory retrieval. Psychol. Rev. 85, 59–108.
Ratcliff, R., Frank, M.J., 2012. Reinforcement-Based Decision Making in Corticostriatal Circuits:Mutual Constraints by Neurocomputational and Diffusion Models, vol. 1229, pp. 1186–1229.
Ratcliff, R., Smith, P.L., 2011. Perceptual discrimination in static and dynamic noise. J. Exp. Psychol. Gen. 139, 70–94. https://doi.org/10.1037/a0018128 (Perceptual).
Ratcliff, R., Smith, P.L., Brown, S.D., McKoon, G., 2016. Diffusion decision model: current issues and history. Trends Cogn. Sci. 20, 260–281. https://doi.org/10.1016/j.tics.2016.01.007.
Ratcliff, R., Tuerlinckx, F., 2002. Estimating parameters of the diffusion model: approaches to dealing with contaminant reaction times and parameter variability. Psychon. Bull. Rev. 9, 438–481.
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Class. Cond. II Curr. Res. Theory 21, 64–99. https://doi.org/10.1101/gr.110528.110.
Rodriguez, C.A., Turner, B.M., McClure, S.M., 2014. Intertemporal choice as discounted value accumulation. PLoS One 9. https://doi.org/10.1371/journal.pone.0090138.
Rodriguez, C.A., Turner, B.M., Van Zandt, T., Mcclure, S.M., 2015. The neural basis of value accumulation in intertemporal choice. Eur. J. Neurosci. 42, 2179–2189. https://doi.org/10.1111/ejn.12997.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275 (80), 1593–1599. https://doi.org/10.1126/science.275.5306.1593.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464. https://doi.org/10.1214/aos/1176344136.
Sewell, D.K., Jach, H.K., Boag, R.J., Van Heer, C.A., 2019. Combining error-driven models of associative learning with evidence accumulation models of decision-making. Psychon. Bull. Rev. https://doi.org/10.3758/s13423-019-01570-4.
Shadlen, M.N., Newsome, W.T., 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. 86, 1916–1936.
Shahar, N., Hauser, T.U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N.S.P.N., Dolan, R.J., 2019. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. PLoS Comput. Biol. 15, 1–25. https://doi.org/10.1371/journal.pcbi.1006803.
Spektor, M.S., Kellen, D., 2018. The relative merit of empirical priors in non-identifiable and sloppy models: applications to models of learning and decision-making: empirical priors. Psychon. Bull. Rev. 25, 2047–2068. https://doi.org/10.3758/s13423-018-1446-5.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Ser. Soc. B Stat. Methodol. 64, 583–639.
Steingroever, H., Wetzels, R., Wagenmakers, E.J., 2014. Absolute performance of reinforcement-learning models for the Iowa Gambling task. Decision 1, 161–183. https://doi.org/10.1037/dec0000005.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: an Introduction, second ed. MIT Press. MIT press, Cambridge, MA.
Ter Braak, C.J.F., 2006. A Markov chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. 16, 239–249. https://doi.org/10.1007/s11222-006-8769-1.
Thura, D., Cisek, P., 2016. Modulation of premotor and primary motor cortical activity during volitional adjustments of speed-accuracy trade-offs. J. Neurosci. 36, 938–956. https://doi.org/10.1523/JNEUROSCI.2230-15.2016.
Tuerlinckx, F., De Boeck, P., 2005. Two interpretations of the discrimination parameter. Psychometrika 70, 629.

Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J., Van Maanen, L., 2017. Approaches to analysis in model-based cognitive neuroscience. J. Math. Psychol. 76, 65–79. https://doi.org/10.1016/j.jmp.2016.01.001.

Turner, B.M., Forstmann, B.U., Steyvers, M., 2019. Joint Models of Neural and Behavioral Data, Computational Approaches to Cognition and Perception. Springer International Publishing. https://doi.org/10.1007/978-3-030-03688-1.

Turner, B.M., Miletić, S., Forstmann, B.U., 2018. Outlook on deep neural networks in computational cognitive neuroscience. Neuroimage 180, 117–118. https://doi.org/10.1016/j.neuroimage.2017.12.078.

Turner, B.M., Palestro, J.J., Miletić, S., Forstmann, B.U., 2019. Advances in techniques for imposing reciprocity in brain-behavior relations. Neurosci. Biobehav. Rev. 102, 327–336. https://doi.org/10.1016/j.neubiorev.2019.04.018.

Turner, B.M., Schley, D.R., Muller, C., Tsetsos, K., 2018. Competing theories of multialternative, multiattribute preferential choice. Psychol. Rev. 125, 329–362. https://doi.org/10.1037/rev0000089.

Turner, B.M., Van Maanen, L., Forstmann, B.U., 2015. Informing cognitive abstractions through neuroimaging:the neural drift diffusion model. Psychol. Rev. 122, 312–336.

Turner, B.M., Wang, T., Merkle, E.C., 2017. Factor analysis linking functions for simultaneously modeling neural and behavioral data. Neuroimage 153, 28–48. https://doi.org/10.1016/j.neuroimage.2017.03.044.

Usher, M., McClelland, J.L., 2001. The time course of perceptual choice: the leaky, competing accumulator model. Psychol. Rev. 108, 550–592.

Van Maanen, L., Brown, S.D., Eichele, T., Wagenmakers, E.-J., Ho, T.C., Serences, J.T., Forstmann, B.U., 2011. Neural correlates of trial-to-trial fluctuations in response caution. J. Neurosci. 31, 17488–17495. https://doi.org/10.1523/JNEUROSCI.2924-11.2011.

Van Maanen, L., Couto, J., Lebreton, M., 2016. Three boundary conditions for computing the fixed-point property in binary mixture data. PLoS One 11, 1–11. https://doi.org/10.1371/journal.pone.0167377.

Van Maanen, L., De Jong, R., Van Rijn, H., 2014. How to assess the existence of competing strategies in cognitive tasks: a primer on the fixed-point property. PLoS One 9. https://doi.org/10.1371/journal.pone.0106113.

Van Ravenzwaaij, D., Oberauer, K., 2009. How to use the diffusion model: parameter recovery of three methods: EZ, fast-dm, and DMAT. J. Math. Psychol. 53, 463–473. https://doi.org/10.1016/j.jmp.2009.09.004.

Voskuilen, C., Ratcliff, R., Smith, P.L., 2016. Comparing fixed and collapsing boundary versions of the diffusion model. J. Math. Psychol. 73, 59–79. https://doi.org/10.1016/j.jmp.2016.04.008.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., Wagenmakers, E.J., 2010. Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. J. Math. Psychol. 54, 14–27. https://doi.org/10.1016/j.jmp.2008.12.001.

Wickelgren, W.a., 1977. Speed-accuracy tradeoff and information processing dynamics. Acta Psychol. 41, 67–85. https://doi.org/10.1016/0001-6918(77)90012-9.

Wiecki, T.V., Poland, J., Frank, M.J., 2015. Model-Based Cognitive Neuroscience Approaches to Computational Psychiatry:Clustering and Classification. https://doi.org/10.1177/2167702614565359.

Worthy, D.A., Pang, B., Byrne, K.A., 2013. Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. Front. Psychol. 4, 1–9. https://doi.org/10.3389/fpsyg.2013.00640.

Yechiam, E., Ert, E., 2007. Evaluating the reliance on past choices in adaptive learning models. J. Math. Psychol. 51, 75–84. https://doi.org/10.1016/j.jmp.2006.11.002.

Yu, A.J., Cohen, J.D., 2009. Sequential effects: superstition or rational behavior?. In: Adv. Neural Inf. Process. Syst. 21 - Proc. 2008 Conf., pp. 1873–1880.

Zhang, S., Huang, C.H., Yu, A.J., 2014. Sequential effects:A Bayesian analysis of prior bias on reaction time and behavioral choice. SAVE Proc. 1844–1849.