Developmental Cognitive Neuroscience 40 (2019) 100733

Contents lists available at ScienceDirect

Developmental Cognitive Neuroscience

journal homepage: www.elsevier.com/locate/dcn

Reinforcement learning across development: What insights can we draw from a decade of research?

Kate Nussenbaum, Catherine A. Hartley *

New York University, United States

ARTICLE INFO

Keywords:
Computational modeling
Reinforcement learning
Decision making

ABSTRACT

The past decade has seen the emergence of the use of reinforcement learning models to study developmental change in value-based learning. It is unclear, however, whether these computational modeling studies, which have employed a wide variety of tasks and model variants, have reached convergent conclusions. In this review, we examine whether the tuning of model parameters that govern different aspects of learning and decision-making processes vary consistently as a function of age, and what neurocognitive developmental changes may account for differences in these parameter estimates across development. We explore whether patterns of developmental change in these estimates are better described by differences in the extent to which individuals adapt their learning processes to the statistics of different environments, or by more static learning biases that emerge across varied contexts. We focus specifically on learning rates and inverse temperature parameter estimates, and find evidence that from childhood to adulthood, individuals become better at optimally weighting recent outcomes during learning across diverse contexts and less exploratory in their value-based decision-making. We provide recommendations for how these two possibilities — and potential alternative accounts — can be tested more directly to build a cohesive body of research that yields greater insight into the development of core learning processes.

1. Introduction

From an early age, people are capable of using past positive and negative experiences to guide their behavior. Young infants can learn to perform actions that elicit reward, like increasing the rate at which they kick their feet to move a mobile hanging overhead (Rovee and Rovee, 1969), or babbling more in response to positive social feedback (Rheingold et al., 1959). Even perseverative errors in the classic A-not-B task (Piaget and Margaret Trans, 1954) can, in part, be attributed to infants' learning of the association between the action of searching in a particular location and the reward of finding a toy (Marcovitch and Zelazo, 1999; Marcovitch et al., 2002). The ability to associate actions with the outcomes they elicit and to use those associations to guide future decisions is evident throughout childhood, adolescence, and adulthood (Raab and Hartley, 2018). Such reinforcement learning processes are proposed to support diverse adaptive functions that transform over the course of development, including the ability to meet homeostatic needs (Keramati and Gutkin, 2014; Moerland et al., 2018), develop social relationships (Jones et al., 2014; Mataric, 1994), and pursue epistemic goals (Oudeyer et al., 2007). Given the centrality of these learning processes to behavior across domains, for years, developmental researchers have asked how acquiring learned associations and deploying them to guide behavior change across the lifespan (Bolenz et al., 2017; DePasque and Galván, 2017).

In the mid-20th century, advances in the application of mathematical models to human associative learning provided researchers with powerful new tools to describe these fundamental processes (Rescorla et al., 1972; Sutton et al., 1998; Witten, 1977). Examining reinforcement learning through the application of formal models offers several advantages. First, quantitative models enable researchers to move beyond broad hypotheses to make highly specific predictions about behavior (van den Bos and Eppinger, 2016). Theories about component processes of learning or patterns of developmental change in these processes can be formalized algorithmically and tested by determining how well they account for observable behavior. Unlike non-model-based measures of decision-making, reinforcement learning models can distinguish processes that contribute to learning the value of different options from processes that translate those value estimates into choices. In this way,

* Corresponding author.
E-mail address: cate@nyu.edu (C.A. Hartley).

https://doi.org/10.1016/j.dcn.2019.100733
Received 5 June 2019; Received in revised form 24 October 2019; Accepted 4 November 2019
Available online 6 November 2019

models can help to illuminate developmental change in cognitive processes or neural representations that are otherwise difficult to tease apart. Second, computational models enable researchers to estimate latent variables that may influence behavior but may not be directly measurable in many experiments (van den Bos et al., 2018). For example, when learning through trial and error, individuals may track the value of different choice options. But probing participants to explicitly report these value estimates may disrupt or change how they learn from experience (Brown and Robertson, 2007; Poldrack and Packard, 2003), and the ability to explicitly monitor and report these estimates may themselves change across development (Kuhn, 2000). By linking individuals' estimates of value to their measurable decisions through a specific, quantitative account, reinforcement learning models enable researchers to mathematically define and access these latent quantities and search for their representation in the brain (Gläscher and O'Doherty, 2010).

For well over thirty years, researchers have simultaneously been investigating the development of the ability to learn from reinforcement and applying mathematical models to more precisely explain and predict learning and decision-making processes. These two streams of research have intersected with increasing frequency in the last decade, with a proliferation of studies applying reinforcement learning models to characterize how children and adolescents use the outcomes of their actions to update their beliefs about their environments and guide their future decisions. Researchers have leveraged reinforcement learning models to investigate questions like how children, adolescents, and adults learn from probabilistic feedback in both stable (Christakou et al., 2013; Davidow et al., 2016; Moutoussis et al., 2018; Palminteri et al., 2016; van den Bos et al., 2012) and dynamic environments (Boehme et al., 2017; Decker et al., 2016; Hauser et al., 2015; Javadi et al., 2014; Potter et al., 2017), update their beliefs based on positive, negative, and neutral feedback (Boehme et al., 2017; Christakou et al., 2013; Jones et al., 2014), and integrate advice or other forms of social influence into their estimates of the value of different actions (Decker et al., 2015; Jones et al., 2014; Rodriguez Buritica et al., 2019). In line with this diversity of conceptual questions, this research has also employed a wide variety of task structures, reward distributions, and model variants.

This broad range of topical foci, experimental paradigms, and analysis approaches has made it challenging to draw broad conclusions from the literature as a whole. What have we discovered about how individuals learn across development from this body of work? We propose that there are two possible patterns of consistent, age-related change in learning mechanisms that developmental modeling studies could reveal. It may be the case that across the different ecological learning problems that reinforcement learning tasks have been designed to emulate, there are developmental changes in the "settings" of learning parameters, such that individuals of different age groups demonstrate consistent biases in how they learn from feedback across varied contexts. For example, there may be developmental shifts in the extent to which individuals weight recent feedback when updating their beliefs or in the extent to which they use their value estimates to influence choice. These static learning settings may affect behavior regardless of the extent to which they promote reward maximization within a given environment. If different developmental stages are indeed associated with particular learning biases, then this should be reflected in consistent age-related increases or decreases in model parameters across a wide variety of studies.

Alternatively, rather than there being developmental shifts in the "settings" of particular learning parameters, development may involve changes in an individual's ability to adapt to the demands of different learning environments. It may be the case that rather than demonstrating particular, static learning biases, individuals vary in the extent to which they adjust their learning strategies based on the statistical properties of a given environment to behave adaptively across diverse contexts. For example, some individuals might weight recent feedback more heavily when their environments are rapidly changing and less heavily when they are relatively stable. From childhood to adulthood, individuals may become better (or worse) at adapting the way they learn, such that their learning strategies more (or less) closely approximate optimal behavior across distinct task contexts. Critically, developmental change in learning adaptivity should not lead to consistent relations between age and parameter estimates across studies, since subtle changes in learning environments may alter the extent to which a given value-updating or choice strategy promotes reward acquisition. Instead, developmental change in learning adaptivity should be reflected in age-related differences in the extent to which behavior is optimal across learning contexts.

To determine the extent to which the literature suggests consistent patterns of change in either learning settings or adaptivity, we examined developmental differences in two parameters meant to capture different aspects of the learning process. We focused on *learning rates*, which determine the extent to which individuals weight recent feedback when updating their estimates of the value of different actions, and *inverse temperatures*, which govern the extent to which individuals select high-valued actions or explore lower-valued alternatives. We find that on the surface, learning rate parameter estimates show no clear developmental trends across studies. However, this apparently inconsistent pattern provides hints that across development, individuals may become better at optimally adapting the extent to which they integrate recent outcomes into their estimates of the value of different actions. Unlike learning rates, across diverse studies, inverse temperature parameter estimates frequently decrease with increasing age, potentially reflecting a consistent bias toward exploration earlier in life. However, clear inference of such a developmental "setting" is hindered by the fact that age-related decreases in inverse temperature estimates can also result from researchers' misspecification of younger individuals' value estimation processes within a reinforcement learning model. In this review, we discuss the neurocognitive mechanisms that may give rise to these changes in learning over developmental time. We also provide suggestions for how future work can more directly test these possible trajectories of developmental change in value estimation and choice processes to better highlight points of convergence across different studies.

1.1. Reinforcement learning tasks

The tasks that have been used to probe reinforcement learning share common features. Typically, participants must make a series of sequential choices between two to four different options, like slot machines or decks of cards. Each option probabilistically delivers positive, negative, or neutral outcomes. The reward probability can remain the same throughout the course of the task (O'Doherty et al., 2004), slowly drift across trials (Daw et al., 2006), or reverse at different points in the experiment (Ghahremani et al., 2010; Li et al., 2011; O'Doherty et al., 2003). Participants are typically instructed to try to select the option that will give them the most reward. For example, in one variant of a classic "two-armed bandit" task, participants make repeated selections between two slot machines or "bandits" — one that gives participants points on 80 % of trials, and one that gives participants points on 20 % of trials. By learning from the outcomes of their choices, participants become increasingly likely to select the better bandit.

1.2. Standard reinforcement learning model

A classic reinforcement learning algorithm assumes that individuals learn by incrementally updating their estimates of the value of taking different actions in different states (Rescorla et al., 1972; Sutton et al., 1998). The extent to which an individual updates her value estimate at each time point is governed by her surprise — the difference between the reward she receives by taking a specific action, and her estimate of the amount of reward she thought she would receive. This difference, the reward prediction error, is then scaled by her learning rate and added to her prior estimate. Formally, this process can be expressed as:

$$Q(s,a)_{t+1} = Q(s,a)_t + \alpha * [r_t - Q(s,a)_t]$$

Where, Q indicates the participant's estimate for the value of taking particular action (a) within a particular state (s) at a particular time (t), r indicates the reward received, andα is a free parameter that represents the participant's learning rate.

1.3. The softmax choice function

Most reinforcement learning models are fit to categorical choice data. Thus, the value estimates computed by a basic reinforcement-learning algorithm, or any variant thereof, must be transformed into choice probabilities. One simple transformation would be to assume that participants choose the option with the highest value on each trial. However, this function is often suboptimal for learning, as the sampling of unexplored choice options or options of uncertain value can promote the discovery of rewarding actions, or verify that the current, highest valued option is still better than competing alternatives, particularly in dynamic or novel environments (Cohen et al., 2007; Daw et al., 2006). Indeed, participants fail to exhibit such a maximization policy in their choice behavior across many different experiments (Herrnstein, 2000). Instead, most models assume that value estimates probabilistically influence choices by applying a softmax function (Daw, 2011):

$$P(a_i|s) = \frac{e^{\beta Q(s,a_i)}}{\sum_k e^{\beta Q(s,a_k)}}$$

Where, α_i is a particular action, k is the number of available actions, and β is the inverse temperature — a free parameter that determines the extent to which value estimates govern actions. When β is high, differences in the value estimates of each potential action have a large effect on choice probabilities, whereas when β is low, choices are "noisier," or related to these value differences to a lesser degree (Fig. 1).

Other choice functions propose different processes for transforming value estimates into choice probabilities. In some algorithms, for example, the uncertainty around each option's estimated value also affects the choice probability computation whereas in others, choice probabilities are assumed to be random — independent of both value and uncertainty — on some proportion of trials (Gershman, 2018). Further variants of the choice function account for biases toward repeating or avoiding previously selected actions, regardless of their outcome history (Daw, 2011).

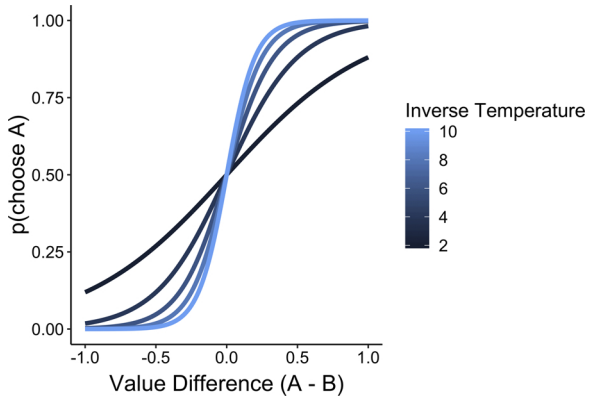

Fig. 1. The softmax function transforms estimates of the value of different options into choice probabilities. The inverse temperature determines the extent to which differences in the value of different options are scaled. When the inverse temperature is high, differences are exaggerated and choices are more deterministic.

1.4. Model comparison

Many studies fit multiple models to each participant's responses to determine which algorithms best capture value-updating and choice processes. Researchers can assess the fit of each model by computing the extent to which it captures participant choices while applying a penalty based on the number of parameters, such that simpler models will be preferred (Aikake, 1974; Stone, 1979). In addition to examining age-related change in parameter estimates, developmental researchers can examine whether individuals at different ages are best fit by different models.

2. What have studies found?

2.1. Learning rates

In the classic formalization of a reinforcement learning model, a single free parameter captures individual differences in the value-updating process. This parameter — the learning rate — reflects the degree to which prediction errors are incorporated into the updated value estimates during learning. High learning rates indicate that individuals weight prediction errors to a greater degree, resulting in value estimates that are heavily influenced by recent outcomes. Low learning rates indicate the opposite: that prediction errors yield small revisions to an individual's value estimate, resulting in choices that are less biased by recent experience and affected by a longer history of choice outcomes. Estimates of learning rates can thus help address the question of how the integration of past experiences to guide actions changes across task contexts, and critically, with age.

At first glance, results from reinforcement learning studies do not suggest any consistent pattern in the tuning of the learning rate parameter across development. Of five developmental studies that reported results from a single-learning rate model, one (Decker et al., 2015) found that learning rates declined with age, with the greatest difference between children (6–12 years old) and adolescents (13–17 years old). Four other studies found either no change in learning rates across adolescents (12–17 years old) and young adults (18–32 years old) (Javadi et al., 2014; Palminteri et al., 2016) or an increase in learning rates across adolescents and adults (Davidow et al., 2016) or across children (8–12 years old), adolescents (13–17 years old), and young adults (18–30 years old) (Master et al., 2019). Few modeling studies have examined how preschool-aged children make value-based decisions, but results from earlier studies of young children's learning strategies that did not leverage formal models are similarly varied — some studies suggest that preschool-aged children (as young as 3) may be more likely to place more weight on recent outcomes relative to older children and adults (Ivan et al., 2018; Schusterman, 1963; Weir, 1964), whereas others suggest that older children weight recent outcomes more heavily than younger children (Berman et al., 1970).

However, this apparent lack of consistency in findings across studies may be better understood by considering an important feature of reinforcement learning: namely, that the optimal setting of a learning rate (the degree to which past versus recent experiences should influence one's current value estimates) depends on the statistics of the particular task or environment (Behrens et al., 2008, 2007; Li et al., 2011; McGuire et al., 2014; Nassar et al., 2012, 2010; O'Reilly et al., 2013). Thus, the heterogeneity of results across developmental studies must be interpreted in light of the heterogeneity of the task structures and reward probabilities that have been used. For example, in environments with deterministic rewards, with the best action leading to reward 100 % of the time, scaling prediction errors to a large degree is optimal, since prediction errors are always indicative of a suboptimal response. In probabilistic environments, however, very high learning rates can cause participants to overweight recent outcomes at the expense of aggregating over a longer history of trials. This can cause recent, rare outcomes — like a negative outcome from a generally good option — to bias

participants' estimates away from the true, underlying expected value of an action. Indeed, two recent studies (Decker et al., 2015; Master et al., 2019) reported opposite patterns of developmental change in learning rates, but they also employed different task structures. In an environment in which the optimal action still led to negative feedback on a small proportion of trials, Decker et al. (2015) found that learning rates decreased from childhood to adulthood, whereas in an environment where the optimal action was always rewarded, Master et al. (2019) found that they increased. Though these two studies reported opposite relations between age and learning rates, when those learning rates are interpreted in the context of each task, both patterns suggest that adults may have better calibrated their learning to the statistics of their environments. For this reason, even across studies applying similar models to address similar questions, developmental findings that appear inconsistent may actually reflect convergent evidence for age-related change in the ability to optimally weight recent feedback across contexts.

Research examining valence-dependent shifts in value updating similarly suggests that development may be marked by improvements in adapting learning to the demands of particular environments. The basic reinforcement-learning algorithm assumes that individuals update their value estimates to the same extent in response to both better-than-expected and worse-than-expected outcomes, but a variant of this model allows for differential scaling of positive and negative prediction errors by introducing separate learning rates for each outcome type (Cazé and van der Meer, 2013; Niv et al., 2012). Across developmental studies using a variety of tasks, model comparison suggests that such two-learning-rate models, which capture valence asymmetries in value updating, provide a better account of participants' learning than single learning-rate models. As with estimates of single learning rates, on the surface, results from two-learning-rate models do not suggest a strong relation between age and parameter tuning. Though two studies found no differences in positive or negative learning rates across age groups (Jones et al., 2014; Moutoussis et al., 2018), three different experiments found age-related decreases in negative learning rates (Hauser et al., 2015; Rodriguez Buritica et al., 2019; van den Bos et al., 2012). This trend corroborates previous behavioral findings (Berman et al., 1970; Ivan et al., 2018; Levinson and Reese, 1967), in which children exhibited a stronger tendency to switch responses after making an incorrect choice than they did to repeat responses after a correct one. Only one study observed an opposite age-related pattern, in which weighting of negative prediction errors increased across adolescence, and weighting of positive prediction errors decreased with age across both adolescents and adults (Christakou et al., 2013).

While valence asymmetries in learning rates might be interpreted as developmental "settings" for the processing of positive and negative outcomes, such asymmetric feedback sensitivity may instead reflect task-specific adaptation of responsivity to signed prediction errors according to what is optimal in that task context. As with single learning rates, the optimal settings of the weighting of positive and negative prediction errors vary depending on the reward statistics of the environment. To illustrate this point, we simulated choice data from agents with two learning rates on two different tasks. The first was a probabilistic selection task similar to that used in van den Bos et al. (2012). Simulated agents completed 150 trials in which they chose between two options — one option rewarded them with a point on 80 % of trials and gave them nothing on 20 % of trials; the other option rewarded them with a point on 20 % of trials and gave them nothing on 80 % of trials. The second task was modeled after the task used in Christakou et al. (2013). Again, simulated agents completed 150 trials in which they chose between two options. One option caused agents to win either 1.9, 2, or 2.1 points on 50 % of trials but lose −2.4, −2.5, or −2.6 points on 50 % of trials; the other option caused them to win either 0.9, 1, or 1.1 points on 50 % of trials but lose −0.4, −.5, or −.6 points on 50 % of trials. For each task, we simulated data from 40,000 agents with positive and negative learning rates that ranged from 0.01 to 1, with a fixed inverse temperature for each environment and calculated the reward earned by each agent.

As Fig. 2 illustrates, the optimal learning rates differ across contexts. In the first case, having relatively equivalent, or slightly positively biased, positive and negative learning rates helps to maximize reward gain, a pattern associated with increasing age across children, adolescents, and adults in van den Bos et al. (2012). In the second case, having a high negative learning rate helps the agent to reduce their value estimate for the option with the high gains but even higher losses, which results in better performance overall. This pattern was observed with increasing age across adolescents and adults in Christakou et al. (2013). Thus, while the settings of positive and negative learning rates in van den Bos et al. (2012) and Christakou et al. (2013) exhibit seemingly opposite developmental patterns, both studies reveal a pattern of more optimal weighting of valenced prediction errors with increasing age.

A few developmental studies have modeled participants' choice behavior in task environments with abrupt changes in reward contingencies (Boehme et al., 2017; Hauser et al., 2015; Javadi et al., 2014). In these types of contexts, *any* static setting of the learning rate may impede performance. Instead, it may be optimal for individuals to increase their learning rates, weighting recent observations more heavily, when increases in prediction error indicate changes in reward probabilities and reduce them during more stable periods (Behrens et al., 2007; Li et al., 2011; Nassar et al., 2012). Only one developmental study has applied a model that allows the learning rate to dynamically shift based on the magnitude of the prediction error (Javadi et al., 2014). Here, adolescents and adults had both similar baseline learning rates and also dynamically adjusted their learning rates in a similar manner. Whether younger individuals exhibit similar adaptation of learning to the dynamics of environmental reward contingencies has not yet been examined.

Understanding the neural mechanisms that might underpin adaptability is challenging for myriad reasons. While many studies have examined developmental differences in neural responsivity to reward and punishment (Silverman et al., 2015; Somerville et al., 2010), fewer studies have focused on responsivity to valenced feedback in the context of learning, when the weight given to different outcomes influences an individual's ability to gain reward. Both cross-sectional and longitudinal developmental studies have observed larger striatal responses to reward (Galvan et al., 2006; Van Leijenhorst et al., 2010; Braams et al., 2015) and to punishment (Galván and McGlennen, 2013) in adolescents compared to children or adults, but it is unclear how these responses relate to differences in outcome weighting during learning. Studies in adults highlight a central role for the striatum and the ventral medial prefrontal cortex in feedback-based learning and value representation (Bartra et al., 2013). Many studies examining representations of predictions errors and their integration into value estimates across development have observed differential age-related patterns of activation in these regions (Christakou et al., 2013; Hauser et al., 2015; van den Bos et al., 2012). While one learning study reported greater striatal response to positive reward prediction errors in adolescents compared to children and adults (Cohen et al., 2010), such age differences in the striatal response to positive prediction errors, or in positive learning rates, were not apparent in other learning studies (Hauser et al., 2015; van den Bos et al., 2012). Another studied associated high negative learning rates in adolescence with increased activity in the anterior insula (Hauser et al., 2015), a region broadly implicated in the processing of aversive outcomes (Büchel et al., 1998; Samanez-Larkin et al., 2008; Simmons et al., 2004). As with behavior, these differences across experiments could reflect a lack of convergence across studies *or* task-specific modulation of learning. Moving forward, research could elucidate the relative contribution of more static learning biases — and their potential neural underpinnings — from adaptation to task environments by examining how the same individuals learn across multiple contexts in which the optimal learning rate, or learning rate asymmetry, varies.

Many of the examples we have highlighted are suggestive of a pattern of increasingly optimal value updating with age. However, it is

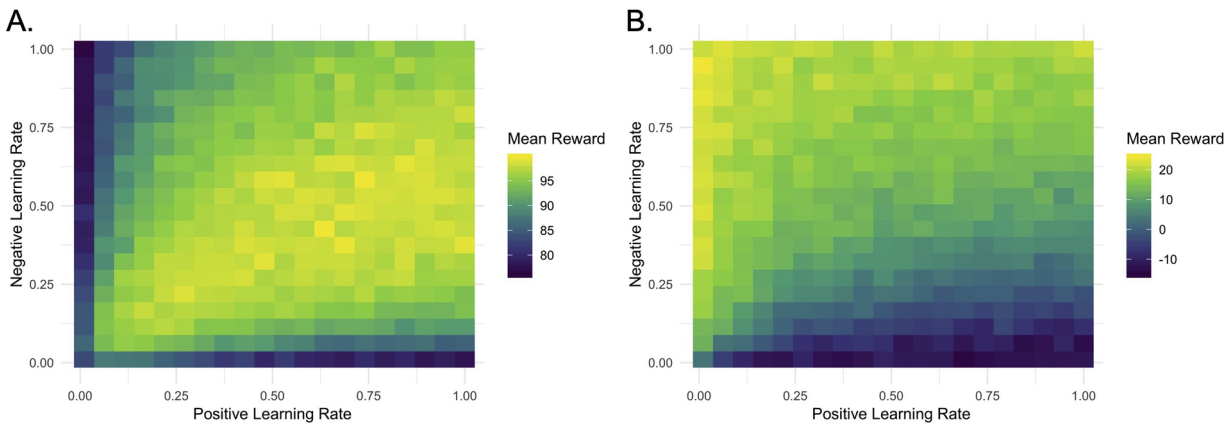

Fig. 2. Simulated data from 40,000 agents in two different learning environments indicate that the optimal asymmetry between positive and negative learning rates differs across contexts. In a two-armed bandit task with static, asymmetric reward probabilities and binary rewards, agents earned the most reward by implementing a slightly higher positive relative to negative learning rate (A). In a two-armed bandit task with static but equivalent reward probabilities, but rewards that differed in their magnitude, agents generally earned the most reward by implementing a very low negative learning rate (B).

difficult to infer such a general, age-related trend across the wide range of tasks and reward contingencies that have been used in previous studies, particularly because many do not report the optimal parameter settings given the reward statistics of the task environment. Behavioral measures of performance partially corroborate the account of age-related improvement in optimal decision-making. Though participants across age groups often demonstrated no differences in their proportion of optimal choices or in the total reward they earned throughout the task (Boehme et al., 2017; Cohen et al., 2010; Hauser et al., 2015; Javadi et al., 2014; van den Bos et al., 2012), in many contexts, older participants made more optimal choices than younger individuals (Christakou et al., 2013; Decker et al., 2015; Moutoussis et al., 2018; Palminteri et al., 2016; Rodriguez Buritica et al., 2018). Only one study found that adolescents outperformed adults (Davidow et al., 2016). This behavioral metric, however, must be interpreted cautiously. Participants' proportions of optimal choices and the total reward they earned throughout the experiment are influenced not just by their integration of reward into value estimates during learning, but also by the way in which these value estimates influence choice, which we discuss in the next section. Reporting how the parameter estimates that capture participants' learning deviate from those of an optimal agent may provide greater insight into the specific components of learning that drive age-related performance differences and would make results easier to compare across different studies. Additionally, the hypothesis that with increasing age, individuals become better at adapting their learning to different contexts could be tested more directly through studies examining whether older individuals' tuning of their learning rates more closely approximates the optimal setting across diverse task environments.

2.2. Inverse temperatures

During learning, individuals must not only estimate the value of different choice options, but also determine how to use those values to guide action selection. The inverse temperature parameter determines the degree to which value estimates influence choice; high inverse temperatures indicate that individuals tend to select the higher-value option while low inverse temperatures indicate that value differences between options govern choices to a lesser extent (See Fig. 1). Estimates of inverse temperatures can help address the question of how the use of learned values to guide decision-making changes across development.

Across reinforcement learning studies, developmental change in the inverse temperature parameter follows a somewhat consistent pattern. All studies we examined either did not report age-related differences in the inverse temperature parameter, found no differences (Davidow et al., 2016; Hauser et al., 2015; Moutoussis et al., 2018; van den Bos et al., 2012), or found that inverse temperatures increased with increasing age (Christakou et al., 2013; Decker et al., 2015; Javadi et al., 2014; Palminteri et al., 2016; Rodriguez Buritica et al., 2019). No studies reported *decreases* in inverse temperatures with age. Suggesting a possible deviation from this general trend, a few early studies found that in prediction tasks, preschool-aged children more consistently selected the more probable outcome relative to adults (Brackbill and Bravos, 1962; Derks and Paclisanu, 1967). However, the majority of studies that have applied reinforcement learning models to developmental data have only included participants over the age of seven, and these model estimates suggest that from middle childhood, increases in age are related to more consistent choices for options that participants have learned yield greater rewards. These findings suggest that, across development, there may be a change in the "setting" of this key learning parameter such that choices are increasingly dictated by their potential to yield reward. But if indeed there is developmental change in the tuning of the inverse temperature parameter, what cognitive mechanisms underlie it?

2.2.1. Exploration

The selection of choice options with non-maximal expected values is often thought of as "exploration" (Daw et al., 2006), and as such, lower inverse temperatures earlier in life have often been attributed to children's greater tendency to explore their environments (Blanco and Sloutsky, 2019; Gopnik et al., 2015, 2017; Sumner et al., 2019). But exploration itself may be driven by multiple, distinct, neurocognitive processes. Studies of exploration have identified at least two forms of exploratory behavior: Individuals may explore either in an uncertainty-directed manner by directing their choices toward options that will reveal the most information, or randomly by making stochastic or "noisy" decisions that are not driven by either reward or information gain (Wilson et al., 2014).

2.2.1.1. Random exploration. Random exploration can be an effective learning strategy across many environments (Bridle, 1990; Sutton et al., 1998; Wilson et al., 2014), and it does not require complex computation, making it a potentially useful strategy for more resource-constrained learners, like children. Indeed, many studies have found that children do demonstrate a bias toward a more stochastic choice policy relative to adults (Blanco and Sloutsky, 2019; Gopnik et al., 2015, 2017; Sumner et al., 2019). For example, Sumner et al. (2019) found that 6-to-12-year-old children tended both to select lower-valued choice options and to switch their responses from trial to trial more frequently than adults in both an environment with static rewards and one with dynamic rewards. Though they earned less reward than adults in both contexts, they were

better at identifying the most rewarding option in the dynamic environment, which may suggest that an early bias toward exploration is an evolutionarily adaptive mechanism that facilitates learning of environmental structure.

Children's and adolescent's "noisier" choices may also be driven by a tendency to avoid repeating the same actions, regardless of how rewarding they are. Some studies have modeled this tendency directly, by including a "perseverative" or "stickiness" parameter that captures a participant's tendency to re-select the previously selected option (Christakou et al., 2013; Decker et al., 2016). These studies found that children and adolescents tended to be less "sticky" than adults, in that they switched responses more often. Early work on children's prediction and decision-making similarly suggests that middle childhood — the developmental stage of the youngest participants typically included in recent reinforcement learning work — may be characterized by an elevated tendency to alternate responses (Craig and Myers, 1963; Gratch, 1959; Ross and Levy, 1958; Weir, 1964). Including a specific model parameter to account for perseveration can distinguish this form of random exploration or stochastic behavior from other mechanisms that implement a non-maximizing choice policy. If studies do not include this parameter, any variance in choice driven by perseverative tendencies will be absorbed by the inverse temperature parameter.

2.2.1.2. Directed exploration. Though children may implement a more stochastic choice policy, their greater exploratory tendencies may also be driven by a stronger desire to resolve uncertainty. Experiments that have directly probed children's learning strategies suggest that early, children often direct their exploration toward the most uncertain parts of the environment (Blanco and Sloutsky, 2019). Many studies have found that children engage more with toys or puzzles when they are less certain about their causal mechanisms (Bonawitz et al., 2012, 2011; Cook et al., 2011; Denison et al., 2013; Gweon et al., 2014; Gweon and Schulz, 2008; Schulz and Bonawitz, 2007; van Schijndel et al., 2015). The inverse temperature parameter used across reinforcement learning studies may reflect, in part, some of this early uncertainty-resolving behavior. As individuals learn which option is most rewarding and select it more often, they also become more certain about its reward probability. In contrast, they select lower-valued options less frequently, and do not accumulate as much experience with their outcomes. Thus the lower-valued options also tend to be less well-known, such that their selection on some trials could in part be explained by individuals acting to reduce their uncertainty to the greatest extent (Wilson et al., 2014). One recent study found that this uncertainty-driven exploration may influence children's performance on value-based learning tasks (Blanco and Sloutsky, 2019). Specifically, relative to adults, 4-year-old children more often selected choice options that they had not recently selected and more often selected those with hidden, as opposed to visually explicit, rewards, though there was extensive heterogeneity in choice behavior within the group. The authors suggest that young children's tendency to distribute their attention may bias them toward behavior that maximizes information-gain, as opposed to choice behavior that maximizes reward. It is important to note, however, that this study only included young children and adults; it remains unclear whether and how a bias toward uncertainty-reduction may interact with value-based learning across middle childhood and adolescence, into adulthood.

2.2.1.3. Strategic exploration. As with learning rates, the optimal tuning of the temperature parameter depends on the structure of the environment. In dynamic contexts with changing reward probabilities (e.g., Decker et al., 2016; Javadi et al., 2014) high rates of exploration throughout the task are necessary to continuously discover newly rewarding options. But in environments in which reward probabilities are static and easy to learn, participants can maximize reward by exploring only minimally to learn which option tends to be more fruitful. Once learners discover the better option, exploration yields no benefit. Thus, while children do seem to explore to a greater extent than adults, it is not clear whether there are also developmental differences in individuals' ability to adapt their rate of exploration to the statistics of their environments.

It may be the case that the most pronounced developmental changes in exploration are changes in *strategic* exploration — or the extent to which individuals adjust their exploratory behavior based on its utility. One study that quantified random, directed, and strategic exploration across 12-to-28-year-olds found evidence only for age-related change in strategic exploration (Somerville et al., 2017). Specifically, older participants increased the frequency with which they selected the more uncertain choice option when they would have more opportunities to use the information they acquired, suggesting their rate of directed exploration was sensitive to its utility. A recent reward-learning study similarly found that while children and adults demonstrated comparable rates of exploration at the beginning of a task, adults more rapidly switched to a strategy of consistently selecting the highest-valued option (Plate et al., 2018). Though many mechanisms could explain this finding, one possibility proposed by the authors is that adults were better than children at strategically reducing their exploration as it became less beneficial. These findings suggest that, though children may demonstrate increased choice stochasticity relative to adolescents and adults, developmental differences in inverse temperature estimates may also sometimes reflect age-related improvement in the ability to strategically adapt rates of exploration to the specific task environment.

2.2.1.4. Neural mechanisms of exploration. Different neurocognitive mechanisms may account for the changes in random, uncertainty-directed, and strategic exploration that may contribute to age-related differences in inverse temperature parameter estimates. The developing brain itself may be "noisier," leading to more variable behavior on some cognitive tasks in childhood and early adolescence (Li et al., 2004; MacDonald et al., 2006). For example, when performing both cognitive control and memory tasks, school-aged children and adolescents demonstrate more intraindividual variability in their response times relative to adults (McIntosh et al., 2008; Tamnes et al., 2012; Williams et al., 2005). The emergence of greater behavioral consistency in early adulthood has been linked to increases in white matter integrity (Tamnes et al., 2012) and, counterintuitively, to increases in the variability of neural activity, which may indicate the flexible engagement of different functional brain networks during demanding cognitive tasks (McIntosh et al., 2008).

While studies have investigated neural mechanisms of behavioral stochasticity across development, less is known about changes in the neural mechanisms that support *directed* exploration. Research in adults has found that the rostrolateral prefrontal cortex (RLPFC), a brain region implicated in relational and analogical reasoning (Bunge, 2004; Bunge et al., 2009) may track the relative uncertainty around different choice options, and promote uncertainty-directed exploration (Badre et al., 2012). Genetic evidence further suggests that dopamine function in the prefrontal cortex may also relate to uncertainty-directed exploration (Frank et al., 2009); individuals with an allele associated with enhanced prefrontal dopamine function demonstrated a greater use of directed exploration. These studies both highlight potential focal points of further investigation of the neural mechanisms that may support changes in directed exploration across development. The RLPFC and its connectivity to other neural regions continues to change throughout adolescence (Wendelken et al., 2016, 2011). However, to the best of our knowledge, no studies have investigated how the development of the RLPFC may relate to changes in exploratory decision-making strategies in reinforcement learning contexts.

Taken together, prior work suggests that multiple different forms of exploration — and unique neural mechanisms — may contribute to observed changes in inverse temperature parameter estimates across development. To better understand the neurocognitive mechanisms of

change in value-based decision-making across development, future work should combine behavioral tasks that can isolate different exploratory strategies with neuroimaging measures that can track their underlying neural instantiations.

2.2.2. Model mismatch

Though inverse temperature parameter estimates can capture exploratory behavior, they may also capture noise in the value estimates themselves, with lower inverse temperatures arising when there is a mismatch between the cognitive learning algorithms implemented by participants and the mathematical learning algorithms implemented by the model (Palminteri et al., 2016; Wyart and Koechlin, 2016). For example, a child may believe that the task is structured such that each choice option will always pay out a loss after it pays out two rewards in a row. In this case, she may always select option B after receiving two rewards in a row from option A. Here, her choices would be highly consistent with her beliefs about the value of the two options, but simple single or dual-learning-rate models would not be able to capture this structured fluctuation of her value estimates. This in turn, would lead the model to suggest that she was frequently selecting the choice she believed was lower value, leading to a low inverse temperature parameter estimate that could be mistaken for "exploration" of the low-value option.

In line with this hypothetical example, Palminteri et al. (2016) found that in a task in which reward probabilities for choice options were anti-correlated, adults' choices were best captured by a model that included a counterfactual learning module that updated value estimates for both the chosen and unchosen option, as well as a contextual module that tracked the average value of sets of choice options to enable similar performance in reward and punishment contexts. Palminteri et al. (2016) found that not only did the most complex model best capture adult choices, but that in adults, inverse temperatures also increased with model complexity. This suggests that the inverse temperature parameters in the simpler models were capturing noise in the model's value estimates. When that "noise" was accounted for by additional modules that more accurately reflected adults' learning strategies, their choices were more deterministically related to the model's value estimates. Palminteri et al. (2016) showed that differences in inverse temperatures may arise due to differences in the extent to which value-updating algorithms accurately capture participants' learning process. We suggest that differences in inverse temperatures *across age groups* may arise for the same reason. To better disambiguate exploration from a mismatch between model algorithms and participant learning strategies, developmental studies should include larger, hypothesis-driven test sets of models that might account for a wide range of learning strategies that individuals might implement at different stages of development.

Of course, even a very large test set of models may not include one that perfectly captures the idiosyncratic learning and decision strategies of every participant. Returning to our previous example, a participant might believe that rewards are distributed with a particular structure (i.e. two rewards followed by a loss), but testing models that represent these plausible, but highly specific, structured beliefs, quickly becomes impractical, if not impossible. A participant might believe that an option pays out a loss after each reward, or every two rewards, or every three rewards, or they might believe that inconsequential features of the task — like the side of the screen on which an option appears — determines her payout. Children and adolescents in particular may approach tasks with a broader set of hypotheses about their structures (Gopnik et al., 2015, 2017). Relative to adults, children's beliefs may be less constrained by prior knowledge, enabling them to more flexibly update their beliefs or conceive of novel solutions to problems (German and Defeyter, 2000; Lucas et al., 2014). The protracted development of the prefrontal cortex may underlie the shift from more flexible, divergent thinking to more constrained and efficient goal-directed behavior (Thompson-Schill et al., 2009). These findings suggest that as a group, children may demonstrate greater heterogeneity in their beliefs about their learning environments, which may inform their estimates of the value of their actions, preventing their choices from being well-captured by any single model. Researchers could better test this idea by examining how the variability in the fit of different models differs across age groups.

3. Conclusions

Over the past ten years, many studies have applied computational models to examine how value-based learning changes with age. Many experiments have applied similar models to address related questions about how learning develops, and when taken together as a whole, their results can appear contradictory. From childhood to adulthood, learning rates increase (Davidow et al., 2016; Master et al., 2019) decrease (Decker et al., 2015), or do not change (Javadi et al., 2014; Palminteri et al., 2016). Inverse temperatures remain constant (Davidow et al., 2016; Hauser et al., 2015; Moutoussis et al., 2018; van den Bos et al., 2012), increase (Christakou et al., 2013; Decker et al., 2015; Javadi et al., 2014; Palminteri et al., 2016; Rodriguez Buritica et al., 2019), or vary within a single context (Plate et al., 2018). But the lack of a simple story regarding the relation between age and the tuning of model parameters is a feature of value-based learning, not a bug. Developmental changes in core learning processes may reflect age-related differences in how individuals adapt their behavior — like the extent to which they weight recent outcomes — to different environments, rather than to more stable settings of parameters. Still, some potential trends in parameter tuning emerged. Inverse temperatures, which reflect the degree to which value estimates influence choices, may increase with age, a pattern that has been attributed to greater stochastic and exploratory behavior earlier in development, but which may be better explained by the developing ability to strategically align one's level of exploration to its utility in a given context. This pattern may also reflect a higher degree of mismatch between the value-updating algorithms implemented in computational models and those that actually control children's and adolescents' learning processes.

In discussing these results, we have laid out a few suggestions for building a more interpretable body of developmental reinforcement learning literature. First, findings may be easier to compare across studies if researchers report not just the parameter estimates that best capture participant data, but also the range of parameter settings, or combinations of parameter settings, that enable optimal performance on the task. Additionally, more studies should test the extent to which participants adjust their tuning of learning parameters when faced with different environmental structures or statistics (Bolenz et al., 2019; Dorfman and Gershman, 2019; Kool et al., 2017). Reporting deviations from optimality and examining adaptability across contexts — both at the behavioral and neural level — will enable researchers to better tease apart the extent to which developmental change in model parameter estimates reflect more stable, context-independent learning biases that are present at particular developmental periods versus differences in the extent to which participants flexibly adapt their learning to the demands of distinct environments.

In reviewing prior work, we have taken reported parameter estimates and model-fitting results at face value, but many studies have not examined the extent to which parameters or tested models are identifiable and recoverable given the specific tasks used. Models that propose different value-updating algorithms, like one- vs. two-learning rate models, may not make sufficiently different predictions in some contexts for them to be differentiated. It may also be the case that different parameter settings would not lead participants to make different choices in some task contexts, preventing parameters from being accurately estimated. Other papers have more extensively laid out recommendations for best practices in model-fitting, and we refer readers to them for a more comprehensive overview of steps that can be taken to better ensure the collection of interpretable data and the accuracy and

robustness of the inferences drawn from computational analyses (Lee et al., 2019.; Wilson and Collins, 2019.). Given the large investments of time and resources required to conduct developmental research, ensuring that we are collecting meaningful data and using the computational tools at our disposal — like simulation — to ensure the robustness of our findings, is particularly important.

In a similar vein, many of the neuroimaging studies that have examined changes in value-based learning have relied on relatively small samples of around 40 participants (Christakou et al., 2013; Cohen et al., 2010; Hauser et al., 2015). While different findings across studies could elucidate how changes in task context trigger the engagement of distinct neural mechanisms, they could also reflect false positive, spurious findings (Button et al., 2013; Poldrack et al., 2017; Turner et al., 2018). The literature could thus benefit from more studies that test not just how behavior varies across different contexts, but also from direct replications of previously reported findings. Finally, though researchers often search for neural correlates of developmental differences in parameter estimates, model-based fMRI may not always be sensitive to such differences (Wilson and Niv, 2015).

The majority of studies that we reviewed relied on cross-sectional samples of school-aged children, adolescents, and adults to measure developmental differences in learning processes. One shortcoming of the literature is that preschool-aged children have been largely excluded from this work. Including very young children in modeling studies is challenging, because reliably estimating reinforcement learning model parameters requires a large amount of data from each participant. Preschoolers may lack the requisite attention spans to complete the number of trials this data-hungry methods requires. Tasks that are "gamified," and those that can be performed repeatedly across different testing sessions, could help to surmount this challenge. Additionally, while cross-sectional studies have illuminated developmental *differences* in learning, they have not provided direct evidence that these differences reflect normative, developmental trajectories. Future work should directly test how learning processes change over time by employing longitudinal designs (Crone and Elzinga, 2015). One challenge with studying learning processes through repeated testing is that it may be difficult to tease apart performance improvements due to participants' increasing exposure to and experience with particular tasks from more general, age-related improvement in learning ability. However, the effects of experience with a particular task likely dissipate with time and might be mitigated if researchers change their task framing and narratives while maintaining the same cognitive demands.

This review focused on simple implementations of reinforcement learning models. Of course, in the real word — and in experimental tasks with greater complexity — many different cognitive processes interact with the basic value-updating and choice mechanisms we discussed here. Studies of adults have leveraged models with more complex value-updating algorithms or task representations to characterize interactions between reinforcement learning and episodic memory (Bornstein and Norman, 2017; Gershman and Daw, 2017), working memory (Collins and Frank, 2009; Collins et al., 2017), attention (Leong et al., 2017; Niv et al., 2015), planning (Daw et al., 2011; Lally et al., 2017), and causal inference (Dorfman et al., 2019; Gershman et al., 2015), yielding insight into how different processes work together to support learning and adaptive action in more complex environments. Several recent empirical studies have extended these lines of inquiry into developmental populations (Cohen et al., 2019; Davidow et al., 2016; Decker et al., 2016; Master et al., 2019.; Moutoussis et al., 2018; Raab and Hartley, 2019). This work holds the potential to better elucidate how known developmental changes in cognitive processes inform reinforcement learning over development. Moreover, by testing samples with a wide age range of participants, in which these cognitive processes will be more variable, developmental research may be able to provide fundamental insights into the nature of the interaction between key component processes of cognition and value-based learning and decision making.

Though modeling reinforcement learning can provide insight into the way in which participants of different ages learn from the outcomes of their actions, it does not inherently provide insight into the processes that govern developmental change itself. As we continue down this road of inquiry, we may reach clearer conclusions about how learning rates, temperatures, or even the value-updating algorithms that individuals implement change with age. But these models are inherently descriptive, not mechanistic. Ultimately, our mandate as researchers who seek to understand development itself is to ask *why* these algorithms, or the settings at which they are implemented, change with the accumulation of experience, the maturation of critical neural functions, or interactions between them.

Acknowledgments

This work was supported by a Klingenstein Simons Fellowship in Neuroscience, a Jacobs Foundation Research Fellowship, a NARSAD Young Investigator Award, and a National Science Foundation CAREER Award Grant No. 1654393 (to C.A.H.), and a National Defense Science and Engineering Graduate Fellowship (to K.N.).

References

Aikake, H., 1974. A new look at the statistical model identification. Inst. Electr. Electron. Eng. Trans. Autom. Control 19 (6), 716–723.
Badre, D., Doll, B.B., Long, N.M., Frank, M.J., 2012. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. Neuron 73 (3), 595–607.
Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage 76 (C), 412–427.
Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S., 2008. Associative learning of social value. Nature 456 (7219), 245–249.
Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10 (9), 1214–1221.
Berman, P.W., Rane, N.G., Bahow, E., 1970. Age changes in children's learning set with win-stay, lose-shift problems. Dev. Psychol. 2 (2), 233.
Blanco, N.J., Sloutsky, V., 2019. Systematic exploration and uncertainty dominate young children's choices. Psyarxiv. https://doi.org/10.31234/osf.io/72sfx.
Boehme, R., Lorenz, R.C., Gleich, T., Romund, L., Pelz, P., Golde, S., et al., 2017. Reversal learning strategy in adolescence is associated with prefrontal cortex activation. Eur. J. Neurosci. 45 (1), 129–137.
Bolenz, F., Kool, W., Reiter, A.M., Eppinger, B., 2019. Metacontrol of decision-making strategies in human aging. eLife 8. https://doi.org/10.7554/eLife.49154.
Bolenz, F., Reiter, A.M.F., Eppinger, B., 2017. Developmental changes in learning: computational mechanisms and social influences. Front. Psychol. 8, 2048.
Bonawitz, E.B., van Schijndel, T.J.P., Friel, D., Schulz, L., 2012. Children balance theories and evidence in exploration, explanation, and learning. Cogn. Psychol. 64 (4), 215–234.
Bonawitz, E.B., Shafto, P., Gweon, H., Goodman, N.D., Spelke, E., Schulz, L., 2011. The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. Cognition 120 (3), 322–330.
Bornstein, A.M., Norman, K.A., 2017. Reinstated episodic context guides sampling-based decisions for reward. Nat. Neurosci. 20 (7), 997–1003.
Braams, B.R., van Duijvenvoorde, A.C.K., Peper, J.S., Crone, EA, 2015. Longitudinal changes in adolescent risk-taking: A comprehensive study of neural responses to rewards, pubertal development, and risk-taking behavior. J. Neurosci. 35 (18), 7226–7238.
Brackbill, Y., Bravos, A., 1962. Supplementary report: the utility of correctly predicting infrequent events. J. Exp. Psychol. 64, 648–649.
Bridle, J.S., 1990. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In: Touretzky, D.S. (Ed.), Advances in Neural Information Processing Systems, vol. 2, pp. 211–217.
Brown, R.M., Robertson, E.M., 2007. Off-line processing: reciprocal interactions between declarative and procedural memories. J. Neurosci. 27 (39), 10468–10475.
Büchel, C., Morris, J., Dolan, R.J., Friston, K.J., 1998. Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron 20 (5), 947–957.
Bunge, S.A., 2004. Analogical reasoning and prefrontal cortex: evidence for separable retrieval and integration mechanisms. Cereb. Cortex 15, 239–249. https://doi.org/10.1093/cercor/bhh126.
Bunge, S.A., Helskog, E.H., Wendelken, C., 2009. Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. NeuroImage 46 (1), 338–342.
Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14 (5), 365–376.
Cazé, R.D., van der Meer, M.A.A., 2013. Adaptive properties of differential learning rates for positive and negative outcomes. Biol. Cybern. 107 (6), 711–719.
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 (11), 1807–1823.

Cohen, A.O., Nussenbaum, K., Dorfman, H., Gershman, S.J., Hartley, C.A., 2019. The rational use of causal inference to guide reinforcement learning changes with age. Psyarxiv. https://doi.org/10.31234/osf.io/j9zuk.
Cohen, J.D., McClure, S.M., Yu, A.J., 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 362 (1481), 933–942.
Cohen, J.R., Asarnow, R.F., Sabb, F.W., Bilder, R.M., Bookheimer, S.Y., Knowlton, B.J., Poldrack, R.A., 2010. A unique adolescent response to reward prediction errors. Nat. Neurosci. 13 (6), 669–671.
Collins, A., Frank, M., 2009. Within and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc. Natl. Acad. Sci. 115 (10), 2502–2507. https://doi.org/10.1101/184812.
Collins, A.G.E., Ciullo, B., Frank, M.J., Badre, D., 2017. Working memory load strengthens reward prediction errors. J. Neurosci. 37 (16), 4332–4342.
Cook, C., Goodman, N.D., Schulz, L.E., 2011. Where science starts: spontaneous experiments in preschoolers' exploratory play. Cognition 120 (3), 341–349.
Craig, G.J., Myers, J.L., 1963. A developmental study of sequential two-choice decision making. Child Dev. 34, 483–493.
Crone, E.A., Elzinga, B.M., 2015. Changing brains: how longitudinal functional magnetic resonance imaging studies can inform us about cognitive and social-affective growth trajectories. Wiley Interdiscip. Rev. Cogn. Sci. 6 (1), 53–63.
Davidow, J.Y., Foerde, K., Galván, A., Shohamy, D., 2016. An upside to reward sensitivity: the Hippocampus Supports enhanced reinforcement learning in adolescence. Neuron 92 (1), 93–99.
Daw, N.D., 2011. Trial-by-trial data analysis using computational models. In: Decis. Making, Affect, and Learning: Attention and Performance XXIII, vol. 23, 1.
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 (6), 1204–1215.
Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J., 2006. Cortical substrates for exploratory decisions in humans. Nature 441 (7095), 876–879.
Decker, J.H., Lourenco, F.S., Doll, B.B., Hartley, C.A., 2015. Experiential reward learning outweighs instruction prior to adulthood. Cogn. Affect. Behav. Neurosci. 15 (2), 310–320.
Decker, J.H., Otto, A.R., Daw, N.D., Hartley, C.A., 2016. From creatures of habit to goal-directed learners: tracking the developmental emergence of model-based reinforcement learning. Psychol. Sci. 27 (6), 848–858.
Denison, S., Bonawitz, E., Gopnik, A., Griffiths, T.L., 2013. Rational variability in children's causal inferences: the sampling Hypothesis. Cognition 126 (2), 285–300.
DePasque, S., Galván, A., 2017. Frontostriatal development and probabilistic reinforcement learning during adolescence. Neurobiol. Learn. Mem. 143, 1–7.
Derks, P.L., Paclisanu, M.I., 1967. Simple strategies in binary prediction by children and adults. J. Exp. Psychol. 73 (2), 278–285.
Dorfman, H.M., Bhui, R., Hughes, B.L., Gershman, S.J., 2019. Causal inference about good and bad outcomes. Psychol. Sci. 30 (4), 516–525.
Dorfman, H.M., Gershman, S.J., 2019. Controllability Governs the Balance Between Pavlovian and Instrumental Action Selection, p. 596577. https://doi.org/10.1101/596577.
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 (8), 1062–1068.
Galván, A., McGlennen, K.M., 2013. Enhanced striatal sensitivity to aversive reinforcement in adolescents versus adults. J. Cogn. Neurosci. 25 (2), 284–296.
Galvan, A., Hare, T.A., Parra, C.E., Penn, J., Voss, H., Glover, G., Casey, B.J., 2006. Earlier development of the accumbens relative to orbitofrontal cortex might underlie risk-taking behavior in adolescence. J. Neurosci. 26 (25), 6885–6892.
German, T.P., Defeyter, M.A., 2000. Immunity to functional fixedness in young children. Psychon. Bull. Rev. 7 (4), 707–712.
Gershman, S.J., 2018. Uncertainty and Exploration, p. 265504. https://doi.org/10.1101/265504.
Gershman, S.J., Daw, N.D., 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu. Rev. Psychol. 68 (1), 101–128.
Gershman, S.J., Norman, K.A., Niv, Y., 2015. Discovering latent causes in reinforcement learning. Curr. Opin. Behav. Sci. 5, 43–50.
Ghahremani, D.G., Monterosso, J., Jentsch, J.D., Bilder, R.M., Poldrack, R.A., 2010. Neural components underlying behavioral flexibility in human reversal learning. Cereb. Cortex 20 (8), 1843–1852.
Gläscher, J.P., O'Doherty, J.P., 2010. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. Wiley Interdiscip. Rev. Cogn. Sci. 1 (4), 501–510.
Gopnik, A., Griffiths, T.L., Lucas, C.G., 2015. When younger learners can Be better (or at least more open-minded) than older ones. Curr. Dir. Psychol. Sci. 24 (2), 87–92.
Gopnik, A., O'Grady, S., Lucas, C.G., Griffiths, T.L., Wente, A., Bridgers, S., et al., 2017. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. Proc. Natl. Acad. Sci. U.S.A. https://doi.org/10.1073/pnas.1700811114.
Gratch, G., 1959. The development of the expectation of the nonindependence of random events in children. Child Dev. 217–227.
Gweon, H., Pelton, H., Konopka, J.A., Schulz, L.E., 2014. Sins of omission: children selectively explore when teachers are under-informative. Cognition 132 (3), 335–341.
Gweon, H., Schulz, L., 2008. Stretching to learn: ambiguous evidence and variability in preschoolers' exploratory play. Proceedings of the 30th Annual Meeting of the Cognitive Science Society 570–574.
Hauser, T.U., Iannaccone, R., Walitza, S., Brandeis, D., Brem, S., 2015. Cognitive flexibility in adolescence: neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development. NeuroImage 104, 347–354.
Herrnstein, R.J., 2000. The Matching Law: Papers in Psychology and Economics. Harvard University Press.
Ivan, V.E., Banks, P.J., Goodfellow, K., Gruber, A.J., 2018. Lose-shift responding in humans is promoted by increased cognitive load. Front. Integr. Neurosci. 12, 9.
Javadi, A.H., Schmidt, D.H.K., Smolka, M.N., 2014. Adolescents adapt more slowly than adults to varying reward contingencies. J. Cogn. Neurosci. 26 (12), 2670–2681.
Jones, R.M., Somerville, L.H., Li, J., Ruberry, E.J., Powers, A., Mehta, N., et al., 2014. Adolescent-specific patterns of behavior and neural activity during social reinforcement learning. Cogn. Affect. Behav. Neurosci. 14 (2), 683–697.
Keramati, M., Gutkin, B., 2014. Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife 3. https://doi.org/10.7554/eLife.04811.
Kool, W., Gershman, S.J., Cushman, F.A., 2017. Cost-benefit arbitration between multiple reinforcement-learning systems. Psychol. Sci. 28 (9), 1321–1333.
Kuhn, D., 2000. Metacognitive development. Curr. Dir. Psychol. Sci. 9 (5), 178–181.
Lally, N., Huys, Q.J.M., Eshel, N., Faulkner, P., Dayan, P., Roiser, J.P., 2017. The neural basis of aversive pavlovian guidance during planning. J. Neurosci. 37 (42), 10215–10229.
Lee, M.D., Criss, A.H., Devezer, B., Donkin, C., Etz, A., Leite, F.P., et al., 2019. Robust modeling in cognitive science. Psyarxiv 2 (3-4), 141–153. https://doi.org/10.31234/osf.io/dmfhk.
Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., Niv, Y., 2017. Dynamic interaction between reinforcement learning and attention in multidimensional environments. Neuron 93 (2), 451–463.
Levinson, B., Reese, H.W., 1967. Patterns of discrimination learning set in preschool children, fifth-graders, college freshmen, and the aged. Monogr. Soc. Res. Child Dev. 32 (7), 1–92.
Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D., 2011. Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. 1–3.
Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., Baltes, P.B., 2004. Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. Psychol. Sci. 15 (3), 155–163.
Lucas, C.G., Bridgers, S., Griffiths, T.L., Gopnik, A., 2014. When children are better (or at least more open-minded) learners than adults: developmental differences in learning the forms of causal relationships. Cognition 131 (2), 284–299.
MacDonald, S.W.S., Nyberg, L., Bäckman, L., 2006. Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. Trends Neurosci. 29 (8), 474–480.
Marcovitch, S., Zelazo, P.D., 1999. The A-not-B error: results from a logistic meta-analysis. Child Dev. 70 (6), 1297–1313.
Marcovitch, S., Zelazo, P.D., Schmuckler, M.A., 2002. The effect of the number of A trials on performance on the A-not-B task. Infancy 3 (4), 519–529.
Master, S.L., Eckstein, M.K., Gotlieb, N., Dahl, R., Wilbrecht, L., Collins, A.G.E., 2019. Distentangling the systems contributing to changes in learning during adolescence. Biorxiv. https://doi.org/10.1101/622860.
Mataric, M.J., 1994. Learning to behave socially. In: Third International Conference on Simulation of Adaptive Behavior, 617, pp. 453–462 pdfs.semanticscholar.org.
McGuire, J.T., Nassar, M.R., Gold, J.I., Kable, J.W., 2014. Functionally dissociable influences on learning rate in a dynamic environment. Neuron 84 (4), 870–881.
McIntosh, A.R., Kovacevic, N., Itier, R.J., 2008. Increased brain signal variability accompanies lower behavioral variability in development. PLoS Comput. Biol. 4 (7), e1000106.
Moerland, T.M., Broekens, J., Jonker, C.M., 2018. Emotion in reinforcement learning agents and robots: a survey. Mach. Learn. 107 (2), 443–480.
Moutoussis, M., Bullmore, E.T., Goodyer, I.M., Fonagy, P., Jones, P.B., Dolan, R.J., et al., 2018. Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. PLoS Comput. Biol. 14 (12), e1006679.
Nassar, M.R., Rumsey, K.M., Wilson, R.C., Parikh, K., Heasly, B., Gold, J.I., 2012. Rational Regulation of Learning Dynamics by Pupil-linked Arousal Systems. Nature Publishing Group, pp. 1–9.
Nassar, M.R., Wilson, R.C., Heasly, B., Gold, J.I., 2010. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J. Neurosci. 30 (37), 12366–12378.
Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., Wilson, R.C., 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. J. Neurosci. 35 (21), 8145–8157.
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 (2), 551–562.
O'Doherty, J., Critchley, H., Deichmann, R., Dolan, R.J., 2003. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. J. Neurosci. 23 (21), 7931–7939.
O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., Dolan, R.J., 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304 (5669), 452–454.
O'Reilly, J.X., Schüffelgen, U., Cuell, S.F., Behrens, T.E.J., Mars, R.B., Rushworth, M.F.S., 2013. Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. Proc. Natl. Acad. Sci. U.S.A. 110 (38), E3660–E3669.
Oudeyer, P.-Y., Kaplan, F., Hafner, V.V., 2007. Intrinsic motivation systems for autonomous mental development. IEEE Trans. Evol. Comput. 11 (2), 265–286.

Palminteri, S., Kilford, E.J., Coricelli, G., Blakemore, S.-J., 2016. The computational development of reinforcement learning during adolescence. PLoS Comput. Biol. 12 (6), e1004953.
Piaget, Jean, Margaret (Trans), Cook, 1954. The Construction of Reality in the Child. Basic Books, New York, NY, US. https://doi.org/10.1037/11168-000.
Plate, R.C., Fulvio, J.M., Shutts, K., Green, C.S., Pollak, S.D., 2018. Probability learning: changes in behavior across time and development. Child Dev. 89 (1), 205–218.
Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., et al., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18 (2), 115–126.
Poldrack, R.A., Packard, M.G., 2003. Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia 41 (3), 245–251.
Potter, T.C.S., Bryce, N.V., Hartley, C.A., 2017. Cognitive components underpinning the development of model-based learning. Dev. Cogn. Neurosci. 25, 272–280.
Raab, H.A., Hartley, C.A., 2018. The Development of Goal-Directed Decision-Making. Retrieved from:. Goal-Directed Decision Making https://www.sciencedirect.com/science/article/pii/B9780128120989000139.
Raab, H., Hartley, C.A., 2019. Adolescents exhibit reduced Pavlovian biases on instrumental learning. Psyarxiv. https://doi.org/10.31234/osf.io/38vgr.
Rescorla, R.A., Wagner, A.R., et al., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. Classical Conditioning II: Curr. Res. Theory 2, 64–99.
Rheingold, H.L., Gewirtz, J.L., Ross, H.W., 1959. Social conditioning of vocalizations in the infant. J. Comp. Physiol. Psychol. 52 (1), 68–73.
Rodriguez Buritica, J.M., Heekeren, H.R., Li, S.-C., Eppinger, B., 2018. Developmental differences in the neural dynamics of observational learning. Neuropsychologia 119, 12–23.
Rodriguez Buritica, J.M., Heekeren, H.R., van den Bos, W., 2019. The computational basis of following advice in adolescents. J. Exp. Child Psychol. 180, 39–54.
Ross, B.M., Levy, N., 1958. Patterned predictions of chance events by children and adults. Psychol. Rep. 4 (1), 87–124.
Rovee, C.K., Rovee, D.T., 1969. Conjugate reinforcement of infant exploratory behavior. J. Exp. Child Psychol. 8 (1), 33–39.
Samanez-Larkin, G.R., Hollon, N.G., Carstensen, L.L., Knutson, B., 2008. Individual differences in insular sensitivity during loss: anticipation predict avoidance learning. Psychol. Sci. 19 (4), 320–323.
Schulz, L.E., Bonawitz, E.B., 2007. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. Dev. Psychol. 43 (4), 1045–1050.
Schusterman, R.J., 1963. The use of strategies in 2-choice behavior of children and chimpanzees. J. Comp. Physiol. Psychol. 56 (1), 96.
Silverman, M.H., Jedd, K., Luciana, M., 2015. Neural networks involved in adolescent reward processing: an activation likelihood estimation meta-analysis of functional neuroimaging studies. NeuroImage 122, 427–439.
Simmons, A., Matthews, S., Stein, M., Paulus, M., 2004. Anticipation of emotionally aversive visual stimuli activates right insula. Neurorepor5 15 (14), 2261–2265.
Somerville, L.H., Jones, R.M., Casey, B.J., 2010. A time of change: behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues. Brain Cogn. 72 (1), 124–133.
Somerville, L.H., Sasse, S.F., Garrad, M.C., Drysdale, A.T., Abi Akar, N., Insel, C., Wilson, R.C., 2017. Charting the expansion of strategic exploratory behavior during adolescence. J. Exp. Psychol. Gen. 146 (2), 155–164.
Stone, M., 1979. Comments on model selection criteria of Akaike and Schwarz. J. R. Stat. Soc. Series B Stat. Methodol. 276–278.
Sumner, E., Li, A.X., Perfors, A., Hayes, B., Navarro, D., Sarnecka, B.W., 2019. The Exploration Advantage: children's instinct to explore allows them to find information that adults miss. Psyarxiv. https://doi.org/10.31234/osf.io/h437v.
Sutton, R.S., Barto, A.G., et al., 1998. Introduction to Reinforcement Learning, Vol. 135. MIT press Cambridge.
Tamnes, C.K., Fjell, A.M., Westlye, L.T., Østby, Y., Walhovd, K.B., 2012. Becoming consistent: developmental reductions in intraindividual variability in reaction time are related to white matter integrity. J. Neurosci.: Off. J. Soc. Neurosci. 32 (3), 972–982.
Thompson-Schill, S.L., Ramscar, M., Chrysikou, E.G., 2009. Cognition without control: when a little frontal lobe goes a long way. Curr. Dir. Psychol. Sci. 18 (5), 259–263.
Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1, 62.
van den Bos, W., Bruckner, R., Nassar, M.R., Mata, R., Eppinger, B., 2018. Computational neuroscience across the lifespan: promises and pitfalls. Dev. Cogn. Neurosci. 33, 42–53.
van den Bos, W., Cohen, M.X., Kahnt, T., Crone, E.A., 2012. Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. Cereb. Cortex 22 (6), 1247–1255.
van den Bos, W., Eppinger, B., 2016. Developing developmental cognitive neuroscience: from agenda setting to hypothesis testing. Dev. Cogn. Neurosci. 17, 138–144.
Van Leijenhorst, L., Gunther Moor, B., Op de Macks, Z.A., Rombouts, S.A., Westenberg, P.M., Crone, E.A., 2010. Adolescent risky decision-making: neurocognitive development of reward and control regions. Neuroimage 51 (1), 345–355.
van Schijndel, T.J.P., Visser, I., van Bers, B.M.C.W., Raijmakers, M.E.J., 2015. Preschoolers perform more informative experiments after observing theory-violating evidence. J. Exp. Child Psychol. 131, 104–119.
Weir, M.W., 1964. Developmental changes in problem-solving strategies. Psychol. Rev. 71, 473–490.
Wendelken, C., Ferrer, E., Whitaker, K.J., Bunge, S.A., 2016. Fronto-parietal network reconfiguration supports the development of reasoning ability. Cereb. Cortex 26 (5), 2178–2190.
Wendelken, C., O'Hare, E.D., Whitaker, K.J., Ferrer, E., Bunge, S.A., 2011. Increased functional selectivity over development in rostrolateral prefrontal cortex. J. Neurosci. 31 (47), 17260–17268.
Williams, B.R., Hultsch, D.F., Strauss, E.H., Hunter, M.A., Tannock, R., 2005. Inconsistency in reaction time across the life span. Neuropsychology 19 (1), 88–96.
Wilson, R.C., Collins, A., 2019. Ten simple rules for the computational modeling of behavioral data. Psyarxiv. https://doi.org/10.31234/osf.io/46mbn.
Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., Cohen, J.D., 2014. Humans use directed and random exploration to solve the explore–exploit dilemma. J. Exp. Psychol. Gen. 143 (6), 2074.
Wilson, R.C., Niv, Y., 2015. Is Model Fitting Necessary for Model-Based fMRI? PLoS Comput. Biol. 11 (6), e1004237.
Witten, I.H., 1977. An adaptive optimal controller for discrete-time Markov environments. Inf. Control. 34 (4), 286–295.
Wyart, V., Koechlin, E., 2016. Choice variability and suboptimality in uncertain environments. Curr. Opin. Behav. Sci. 11, 109–115.