The Evolution of Cognitive Biases in Human Learning

Peter S. Park[*]

[*]Department of Mathematics, Harvard University,
1 Oxford St., Cambridge, MA 02138, USA
Email: pspark@math.harvard.edu.

January 24, 2022

Abstract

Cognitive biases like underinference, the hard-easy effect, and recurrently non-monotonic confidence are evolutionarily puzzling when viewed as persistent flaws in how people learn from environmental feedback. To explain these empirically robust cognitive biases from an evolutionary perspective, we propose a model of ancestral human learning based on the cultural-evolutionary-theoretic hypothesis that the primary selection pressure acting on ancestral human cognition pertained not to learning individually from environmental feedback, but to socially learning task-specific knowledge. In our model—which is inspired by classical Bayesian models—an ancestral human learner (the student) attempts to learn task-specific knowledge from a role model, with the option of switching between different tasks and role models. Suppose that the student's method of learning from their role model is *a priori* uncertain—in that it can either be successful imitation learning or *de facto* innovation learning—and the ecological fitness costs of meaningfully retaining environmental feedback are high. Then, the student's fitness-maximizing strategy does not retain their environmental feedback and—depending on the choice of model parameters—can be characterized by all of the aforementioned cognitive biases. Specifically, in order for the evolutionarily optimal estimate of confidence in this learning environment to be recurrently non-monotonic, it is necessary (as long as the environment's marginal payoff function satisfies a plausible quantitative condition) that a positive proportion of ancestral humans' attempted imitation learning was unknowingly implemented as *de facto* innovation learning. Moreover, an ecologically rational strategy of selective social learning can plausibly cause the evolutionarily optimal estimate of confidence to be recurrently non-monotonic in the empirically documented way: general increase with an intermediate period of decrease.

Keywords. cognitive biases; overconfidence; underinference; hard-easy effect; non-monotonic confidence; decision theory; causal inference; Bayesian updating models; cultural evolution; selective social learning; ecological rationality

1 Introduction

Humans have evolved to meaningfully incorporate into their beliefs the low-variance, essentially deterministic environmental feedback they observe—the domain of causal inference—so as to improve future decisions (Pinker, 2010). For example, people often learn to pay credit card bills (Agarwal et al., 2008) and return rented videos (Haselhuhn et al., 2012) on time after first paying late fees. However, the same cannot be said when the variance is high. In the domain of high-variance environmental feedback, unbiased Bayesian updating should in theory be normatively rational (Corner & Hahn, 2012) and even evolutionarily optimal (McNamara & Houston, 1980) in many settings. In line with this, a review of 11 empirical studies of animal foraging and reproductive decisions—spanning eight species of birds, three of non-human mammals, one of fish, and one of insects—found the behavior of all but one of the species to be consistent with the predictions of Bayesian updating models (Valone, 2006). For humans, however, learning in settings of high-variance environmental feedback deviates from Bayesian updating in various ways (e.g., Tversky & Kahneman, 1974). These deviations, referred to in the literature as cognitive biases, result from evolved tendencies by which humans systematically fail to learn meaningfully from high-variance environmental feedback.

A myriad of cognitive biases are apparent from the insightful experiments of Sanchez and Dunning (2018, 2020) on human learning. In each variant of their experiment, subjects learned a new task possessing a payoff structure with fixed uncertainty: classifying profiles with lists of properties (for example, symptoms) into categories (for example, made-up diseases). The subjects attempted this task 60 times while simultaneously reporting their confidence: their self-estimate of the probability that their answer is correct. After each of their 60 answers, they received immediate feedback. Despite this, the subjects did not learn from their environmental feedback in a Bayesian-rational manner, as one can see from the following patterns in the data (see Sanchez & Dunning, 2018, Figures 1–4; Sanchez & Dunning, 2020, Figures 1–3).

1. **The subjects' confidence graph—that of their average self-estimate as a function of trial number—was non-monotonic.** Specifically, the confidence graph was comprised

of three phases: a beginning phase of increase, an intermediate phase of decrease, and a final phase that returned to increase. This pattern agrees with the finding of the well-known experiment of Kruger and Dunning (1999) on confidence as a function of true ability—as well as its replications—that the former variable can be a non-monotonic function of the latter (see Burson et al., 2006, Figures 4–6; Haun et al., 2000, Figures 5–7; and Kruger & Dunning, 1999, Figures 2–3). This also agrees with the work of Hoffman and Burks (2020) investigating truckers' self-estimates of the number of miles driven each week, which found their average to be non-monotonic with respect to the level of experience and the average of the true value, monotonically increasing in the level of experience (see Hoffman & Burks, 2020, Figure 1).

2. **The average difference between confidence and the environmental feedback eventually became positive—signifying overconfidence—and proceeded to increase instead of decaying to zero.** This pattern is consistent with the extensive evidence on overconfidence in the cognitive bias literature: for example, as a cause of wars (Dixon, 1976; Johnson, 2004), stock market bubbles (Akerlof & Shiller, 2009; Scheinkman & Xiong, 2003), and underpreparation for catastrophes (MacKenzie, 1994; Schlosser, 2013). Consistently becoming overconfident compared to the environmental feedback, by itself, likely suffices to contradict Bayesian rationality (Augenblick & Rabin, 2021).

3. **The confidence graphs from all variants of the experiment were essentially indistinguishable from each other, even though the subjects of each experimental variant on average performed differently and thus received different environmental feedback.** The confidence graph in essence only depended on the number of past observations, the level of experience. This pattern is consistent with two well-documented cognitive biases: underinference (Benjamin, 2019), the tendency to insufficiently update one's belief in the direction of new evidence compared to Bayesian inference; and the hard-easy effect (Lichtenstein & Fischhoff, 1977; Moore & Healy, 2008), the tendency to be overconfident on difficult tasks and underconfident on easy tasks. Indeed, a predetermined confidence function—one that depends not on past environmental feedback, but only on other types of information like one's level of experience—would generically differ from the Bayesian aggregate of the past environmental feedback. The difference between the two would generically persist, manifesting as both underinference and—depending on the hard-easy effect—either persistent overconfidence or underconfidence.

These three non-Bayesian patterns robustly replicated in all six variants of the Sanchez–Dunning experiment (2018, 2020), including the variant that used the incentive-compatible Becker–DeGroot–Marschak method (Becker et al., 1964) to monetarily incentivize accurate answers. The non-Bayesian inaccuracy of subjects' learning (Jansen et al., 2021) and the persistence of this inaccuracy in the face of monetary incentivization (Ehrlinger et al., 2008) have also been documented in replications of the Kruger–Dunning experiment; these phenomena have been found in the aforementioned work of Hoffman and Burks (2020) on truckers' self-estimates of productivity, as well. Note that the Kruger–Dunning experiment is similar in objective and design to the Sanchez–Dunning experiment. A crucial difference, however, is that accurate environmental feedback is immediately provided by the experimenter in the latter, but not in the former. The Sanchez–Dunning experiment thus compellingly raises the question of why humans have evolved to underinfer from freely available environmental feedback, even when meaningfully learning from it is made easy and monetarily advantageous.

How did our evolutionary past select for cognitive biases, traits that systematically cause errors in judgement? To solve this puzzle, we appeal to cultural evolutionary theory's extensive body of evidence that humans primarily rely on learning from their fellow group members, rather than from the environmental feedback itself (Boyd & Richerson, 1985, 1988, 1995; Cavalli-Sforza & Feldman, 1981; Lew-Levy et al., 2017). This evidence informs and is informed by a central hypothesis of cultural evolutionary theory: that adaptive, socially exchanged, and intergenerationally accumulated knowledge—relevant to fitness-relevant tasks like foraging, reproduction, and warfare—comprised the primary selection pressure acting on ancestral human cognition (Baimel et al., 2021; Henrich, 2015; Humphrey, 1976; Laland, 2017; Muthukrishna & Henrich, 2016; Muthukrishna et al., 2018; Reader et al., 2011; Street et al., 2017; van Schaik & Burkart, 2011; Whiten & van Schaik, 2007).

In this paper, we construct an evolutionary model of human learning based on this cultural-evolutionary-theoretic hypothesis: one in which an ancestral human learns primarily via knowledge learned from group members, rather than via environmental feedback. The model is constructed by modifying a classical Bayesian model of repeated task-learning to veridically represent the hypothesized setting of social, knowledge-based task-learning. Another key modification we add is our assumption that the cognitively constrained agent of our model—representing an ancestral human learner—faces selection pressures against meaningful retention of high-variance environmental feedback, due to onerous ecological fitness costs of overcommitting attention (e.g., increased risks from ambushes and accidental injury caused by a lack of situational awareness). It follows from this assumption that the confidence function comprising the agent's fitness-maximizing strategy is characterized by discrete confidence levels and systematic deviations from classical Bayesian inference (i.e., from unbiased incorporation of environmental feedback), consistent with the empirical finding of Lisi et al. (2021). Specifically, this confidence function is characterized by various cognitive biases like underinference, the hard-easy effect, and—depending on the parameters of our model—recurrent non-monotonicity.

We begin by describing in Subsection 2.1 a finite-outcome-space version of the classical Bayesian decision-theoretic model. This general model serves both as an inspiration for our evolutionary model and as a *reductio ad absurdum* argument that humans may not learn from high-variance environmental feedback via classical Bayesian inference. The contradiction is as follows. Classical Bayesian inference is effective because a Bayesian-updating prior (that has not *a priori* ruled out any possibility) is almost surely guaranteed to eventually converge to the truth: the property of consistency. However, this property is in contradiction with the aforementioned findings from the cognitive biases literature: first, that a human learner's prior (such as that of their ability) can persistently deviate from their past observations; and second, that it can be recurrently non-monotonic with respect to the number of observations, regardless of the actual observations themselves.

We then resolve these empirical contradictions by presenting in Subsection 2.2 our evolutionary model: a modification of the classical Bayesian model, adapted to represent the knowledge-based learning environment of ancestral humans in the context of high-variance payoff observations. In our modified Bayesian model, the agent learns a task over repeated attempts, each of which generates a payoff. When the expected cost of retaining high-variance payoff observations—due to onerous ecological fitness costs from overcommitting attention—is sufficiently high, the agent's optimal learning strategy does not update their prior of their payoff-acquisition ability in the given task (confidence) with respect to the payoff observations. Instead, the agent updates their confidence as a function of information in the complement of payoff observations: in our model, knowledge and the speed of learning. The consequent unavailability of payoff data—the key departure from classical Bayesian decision theory—generates our first desired conclusion: that evolved confidence generically deviates from the past payoff observations in a recurrent manner. This conclusion is a special case of a more general phenomenon: a given learning strategy's systematic departure from classical Bayesian updating when the ancestral learning environment for which it is ecologically rational differs from the contemporary learning environment in which it actually operates (Gigerenzer, 2000; Gigerenzer & Todd, 1999;

McKay & Efferson, 2010). Persistent underinference and the hard-easy effect follow from the recurrent nature of this evolutionary optimal confidence function.

The second desired conclusion—that this recurrent, evolutionarily optimal confidence function can be non-monotonic—follows from incorporating the cultural-evolutionary-theoretic hypothesis that the agent's learning occurs via attempted imitation of a role model. This non-monotonicity can occur due to a dichotomy between successful imitation learning and *de facto* innovation learning: two learning methods whose classification is *a priori* uncertain to the agent.

The details of this dichotomy and of other aspects of our model are presented in Section 2. The predictions of this model are then made mathematical precise in the theorem statements presented in Section 3. The proofs of the theorems can be found in the Appendix.

We thus find that several classes of cognitive biases can be parsimoniously explained as evolutionary byproducts of the idiosyncratically knowledge-based and social nature of ancestral humans' hypothesized learning environment. Often thought of as structural flaws in humans' individual learning, cognitive biases may instead be evolutionarily rooted in two hypothesized characteristics of our ancestral environment: first, the primarily knowledge-based and social—not individual—nature of human learning in natural settings, as theorized by cultural evolutionary theory; and second, ecological fitness costs of meaningfully retaining environmental feedback—due to cognitive constraints—and the consequent pressure to rely instead on setting-specific sources of information, as theorized by the ecological rationality hypothesis (Gigerenzer, 2000; Gigerenzer & Todd, 1999).

2 The model

2.1 Classical Bayesian model

Suppose that an agent repeatedly attempts a task. Each yields a random payoff that is contained in a finite set of values $S \subset \mathbb{R}$. The finiteness of S constitutes the realistic assumption that the agent, due to cognitive constraints, categorizes observations into finitely many bins. The payoff from each task attempt is drawn i.i.d. from a fixed probability distribution $\phi \in \Phi \subseteq \mathcal{P}(S)$, which would depend on the agent's ability to acquire payoffs, the abundance of the environment, and various other factors. Here, $\mathcal{P}(S)$ denotes the set (which can be thought of as a state space) of all probability distributions on S, and $\Phi \subseteq \mathcal{P}(S)$ denotes the subset of probability distributions that may feasibly occur in a given setting.

For the purpose of maximizing payoff, the agent is incentivized to accurately predict the expected value of the future task attempt's payoff. This was likely the case for ancestral human foragers, who by default engaged repeatedly in a highly specialized foraging role (Hooper et al., 2015), but also faced incentives to be opportunistic: to accurately appraise—and based on the result of said appraisal, possibly procure—additional foraging opportunities as they arise (Bird-David, 1992). We model this dichotomy as follows. We assume that before each task attempt, the agent has the choice of forgoing a fraction r of the time spent on it (corresponding to the same fraction of the task attempt's entire payoff) for a payoff whose value is observed beforehand. The opportunity-cost payoff is rc, where c drawn from a fixed distribution $\psi \in \mathcal{P}(S)$ whose support is all of S. It follows that the agent maximizes the immediate payoff by taking the payoff from the task attempt if its mean $r\mathbb{E}[\phi]$ is greater than rc, take the opportunity cost if $r\mathbb{E}[\phi]$ is less than rc, and take either option when $r\mathbb{E}[\phi]$ is equal to rc.

The agent thus benefits from accurately estimating the task attempt's expected payoff $\mathbb{E}[\phi]$. This can likely be achieved by a small number of observations—even just one—when ϕ has low variance. Under our assumption that payoffs are observationally categorized by the agent into finitely many bins, assuming further that the payoffs have low variance amounts to the condition that nearly all payoffs (i.e., close to probability one) fall in a single bin $s \in S$. Consequently, the agent can productively use causal inference, in the sense that assuming every future task attempt will yield the previously observed payoff of s will nearly always be correct. The payoff-maximizing strategy is to choose the higher value between the task attempt's expected payoff $r\mathbb{E}[\phi] \approx rs$; and the observed opportunity cost rc.

The discernment of the payoff distribution ϕ—and more specifically, its expected value $\mathbb{E}[\phi]$—is more difficult when ϕ has high variance. In this domain, more than one bin in S occurs with significant probability. Consequently, the agent will in general need to learn from a large sample size of payoffs in order to asymptotically determine the true state ϕ from the set of *a priori* possible states Φ.

Suppose that the true state ϕ is initially drawn from a probability distribution $\xi \in \mathcal{P}(\Phi)$. Then, Bayes' theorem states that the probability distribution of ϕ conditional on the previous payoff observations being $s_1, s_2, \ldots, s_n$ is given by

$$\xi_{s_1,\ldots,s_n} = \mathcal{B}_{s_n} \circ \cdots \circ \mathcal{B}_{s_2} \circ \mathcal{B}_{s_1}(\xi), \tag{1}$$

where $\mathcal{B}_x : \mathcal{P}(\Phi) \to \mathcal{P}(\Phi)$ is the Bayes'-rule map

$$\mathcal{B}_x(\omega)(\theta) = \frac{\theta(x)\omega(\theta)}{\displaystyle\int_{\hat{\theta}\in\Phi} \hat{\theta}(x)\omega(\hat{\theta})d\hat{\theta}}. \tag{2}$$

Consequently, the payoff-maximizing choice of whether to forgo part of the task-attempt payoff is to compare its expected value

$$r \int_{\phi\in\Phi} \mathbb{E}[\phi]\, d\xi_{s_1,\ldots,s_n}(\phi) \tag{3}$$

with the observed opportunity cost rc. In summary, the agent's evolutionarily optimal strategy overall is to begin with the prior ξ, update it via the Bayes' rule map $\mathcal{B}_s$ in terms of each task attempt's observed payoff s, and decide whether to forgo part of the nth task attempt for an observed opportunity cost by using the prior $\xi_{s_1,\ldots,s_{n-1}}$ at that point in time.

Bayesian inference can be effective even without explicit knowledge of the true distribution ξ from which the state ϕ is drawn. An obvious obstruction to this effectiveness is Cromwell's rule: if a state is not contained in the support of the prior ω, then this will persist in $\omega_{s_1,\ldots,s_n}$ for any sequence of observations $s_1, \ldots, s_n$. It turns out that Cromwell's rule is the only such obstruction when the outcome space S is finite. Specifically, suppose that the true state ϕ is contained in the support of the prior ω. Then, as $n \to \infty$, the nth Bayesian update of ω

$$\omega_n = \omega_{s_1,\ldots,s_n} \tag{4}$$

will converge to the one-point distribution

$$\chi_\phi(\theta) = \begin{cases} 1 & \text{if } \theta = \phi \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

with prior probability one: the property of consistency (Doob, 1949; Freedman, 1963). In other words, even an agent with a misspecified initial prior—for example, one that evolved in a past environment with a different distribution of ϕ—will in all likelihood eventually converge to the true state ϕ, as long as the initial prior is not too restrictive.

The property of consistency yields a practical test to reject the null hypothesis that a given learner is Bayesian in the classical sense. We can do so if the learner's prior does not converge to the (one-point distribution on the) true state as the number of observations goes to infinity. A special case of this test is provided by checking whether a learner's estimate of their expected payoff-acquisition ability converges to the true expected payoff. Indeed, suppose that the learner's prior were updated via classical Bayesian inference while starting from an initial prior ω that has not ruled out the true state ϕ. Then, with prior probability one, the learner's estimate of their expected payoff-acquisition ability

$$\int_{\hat{\theta}\in\Theta} \mathbb{E}[\hat{\theta}]\, d\omega_n(\hat{\theta}), \tag{6}$$

would converge to the true expected payoff

$$\mathbb{E}[\phi] \tag{7}$$

as the number of observations n goes to infinity. While the true expected payoff (7) is unobservable, it will with probability one coincide with the mean of the past payoff data

$$\frac{s_1 + \cdots + s_n}{n} \tag{8}$$

as $n \to \infty$, due to the law of large numbers. We should thus be skeptical of a learner's Bayesianness if their estimate (6) of their expected payoff-acquisition ability does not appear to converge to the mean of the past payoff data (8). Note that this practical test for falsifying a learner's Bayesianness is not new; it is essentially a corollary of standard Bayesian statistics.

To illustrate, consider a gambler who, over repeated attempts, continues to be mistaken about the expected value of a fixed probabilistic lottery. They may persistently believe that the expected payoff from betting their money on a negative-expected-value lottery is positive, even after gambling on it a large number of times while observing the resulting payoff data. Then, we can be reasonably certain that the gambler is not, in the classical sense, Bayesian-updating with respect to their payoff data. We hypothesize that the persistent deviation of the gambler's prior from the true state is caused by the high variance of the payoff data. Other learners who may fail our test for classical Bayesianness include professionals whose priors of their performance persistently deviate from the true value (Park & Santos-Pinto, 2010; Hoffman & Burks, 2020), traders and managers who persistently overestimate future returns on their investments (Barber & Odean, 2001; Malmendier & Tate, 2005), and gymgoers who repeatedly overpay on membership fees based on persistently overoptimistic priors of their attendance rate (DellaVigna & Malmendier, 2006). Such field evidence against the hypothesis that human learning from high-variance payoff data is classically Bayesian corroborates the extensive lab evidence of the relevant cognitive biases.

2.2 Evolutionary model of ancestral human learning

To resolve the predictive inadequacies of the classical Bayesian paradigm, we modify it in the following way. We assume that the agent estimates their payoff-acquisition ability as a function of task-specific knowledge, and not necessarily of the previously observed payoff data. Our evolutionary model incorporates two veridical sources of uncertainty which are sufficient to generate recurrent non-monotonicity. First, tasks vary in difficulty, a value that represents the total amount of knowledge required to completely learn the task. The agent's marginal payoff is a bivariate function of the difficulty value and their current level of knowledge: the subset of the total knowledge they have learned so far.

Second, tasks vary in the method used to learn the relevant knowledge: imitation and innovation. We incorporate into our model the cultural-evolutionary-theoretic finding that the primary source of an ancestral human's task-specific knowledge was learning from role models who were ostensibly proficient in the task—imitation—rather than learning individually from environmental feedback—innovation (Boyd & Richerson, 1985; Cavalli-Sforza & Feldman, 1981). The superior efficiency of imitation learning, especially in the context of intergenerational knowledge accumulation, is hypothesized to have enabled humans' unprecedented evolutionary success.

The dichotomy between imitation learning and innovation learning is confusing at first glance, given that in our model, the student always attempts to imitate a role model. This dichotomy occurs because the helpfulness of role models in providing a genuinely new path forward via imitation learning is not guaranteed. A student may successfully learn via imitation of their role model, as planned. It is also possible that the role model's ostensible proficiency in the task does not translate to productive imitation learning, in which case the student learns by *de facto* innovation. Specifically, the role model may not actually be providing a new learning path that the student would not have accessed if they were to instead learn by innovation. In the context of direct teaching, for instance, this may be due either to the method of teaching (a teacher may use an open-ended or ambiguous teaching method, such as the Socratic method, without actually guiding students to think in a new way) or to the teacher's own limitations (which may not be discernible to students when their environmental feedback has high variance). It would be difficult for the student to deduce from high-variance environmental feedback whether their role model is meaningfully providing them with a new learning path to imitate.

Throughout this paper, the term "task" will denote a student's package comprised of a repeated knowledge-intensive task that produces fitness-aiding payoffs (i.e., foraging for food), their choice of role model for it, and the learning method by which the student obtains the relevant knowledge: classified into imitation learning and innovation learning. The student's task package can be thought of as a pair (j, a) for the type of learning method $j \in \{im, in\}$ with which the student learns the task from the teacher (where $j = im$ denotes imitation and $j = in$ denotes innovation) and the difficulty value $a \in (0, \infty) \cup \{\infty\}$ of the task.

The difficulty value $a \in (0, \infty) \cup \{\infty\}$ of a task denotes the amount a of knowledge the student needs to completely learn it, given the specifics of the task package (the teacher, the learning method, and the task itself). A task with the difficulty value $a = \infty$ represents an impossible one, in that the specifics of the task prevent the student from learning it to completion. Suppose the student currently knows $b \leq a$ of the total amount of knowledge required to completely learn the task. The values of b and a determine the marginal payoff $f(a, b)$, which we assume is strictly increasing in b, strictly decreasing in a, and continuously differentiable. By scaling the marginal payoff values to have minimum 0 and maximum 1, we can suppose that the function $f(a, \cdot)$ maps the domain $[0, a]$ to the range $[0, 1]$. We assume that completely learning a task guarantees the maximum marginal payoff: $f(a, a) = 1$ for every a. Moreover, we assume that impossible tasks—unable to be meaningfully learned—always yield the minimum marginal payoff: $f(\infty, b) = 0$ for all b.

One example of a marginal payoff function

$$f : \{(a, b) \in ((0, \infty) \cup \{\infty\}) \times [0, \infty) : b \leq a\} \to [0, 1] \tag{9}$$

satisfying these conditions is

$$f(a, b) = \left(\frac{b}{a}\right)^{\lambda}, \tag{10}$$

for $\lambda > 0$, which is extended to the point at infinity $a = \infty$ as

$$f(\infty, b) = \lim_{a \to \infty} \left(\frac{b}{a}\right)^{\lambda} = 0. \tag{11}$$

This family of functions is characterized by polynomial growth in b. Another example of such a marginal payoff function is

$$f(a, b) = \zeta^{a-b} \tag{12}$$

for $\zeta \in (0, 1)$, which is also extended to the point at infinity $a = \infty$ as

$$f(\infty, b) = \lim_{a \to \infty} \zeta^{a-b} = 0. \tag{13}$$

This family of functions is characterized by exponential growth in b.

We assume that the risk of an infinitely difficult task $a = \infty$ only exists when $j = in$. In the other case of $j = im$, the learnability of the given task is guaranteed by the teacher already having learned it completely. However, when $j = in$, the teacher may not have actually learned the task completely despite serving as the student's role model. The lack of guarantee of the given task's learnability leads to a nontrivial probability of an unfortunate setting: one in which the student squanders time on attempting to learn an impossible task from a teacher, one or both of whom have not yet realized the said impossibility. The exclusivity of unlearnability to innovation learning can be seen by the comparison between solving an exam problem and solving a research problem. The former—imitation learning—is guaranteed to complete in finite time, because the teacher has solved the problem before assigning it as an exam question. However, the latter—innovation learning—is not guaranteed to complete in finite time. Indeed, a research problem, by definition, is one that has not yet been solved by anyone, so it may *a priori* be impossible to solve. Overall, we assume that the difficulty values of tasks with learning method $j = im$ are distributed as a regular exponential distribution (i.e., with p.d.f. $\mu_{im}(a) = \eta^a \log \frac{1}{\eta}$ for finite a and $\mu_{im}(\infty) = 0$, where $0 < \eta < 1$), whereas the distribution of difficulty values of tasks with learning method $j = in$ is assumed instead to have a positive probability p on $a = \infty$ (i.e., with p.d.f. $\mu_{in}(a) = (1 - p)\eta^a \log \frac{1}{\eta}$ for finite a and $\mu_{in}(\infty) = p$). The overall distribution of tasks (j, a) on

$$\mathcal{U} = \{im, in\} \times ((0, \infty) \cup \{\infty\}), \tag{14}$$

defined by the p.d.f.

$$\mu(j, a) = \begin{cases} q\mu_{im}(a), & \text{if } j = im, \\ (1 - q)\mu_{in}(a) & \text{if } j = in; \end{cases} \tag{15}$$

places probability q on the task's learning type being imitation and $1-q$ on that being innovation.

Other than the risk of unlearnability, the second way in which tasks of learning method $j = im$ differ from those of learning method $j = in$ is in the speed of learning. Regardless of the learning method, the student learns knowledge in discrete jumps, each following a task attempt. Let $B(t)$ denote the knowledge level after the tth task attempt, where $B(0) = 0$, meaning that the initially naive student has knowledge $b = B(0) = 0$ of the task when starting out. The discrete knowledge levels $0 = B(0) < B(1) < \cdots$ are assumed to satisfy $\lim_{t\to\infty} B(t) = \infty$. The amount of time the tth task attempt takes for the student is assumed to differ between the two learning types. Let $\Delta_{im}(t)$ (respectively, $\Delta_{in}(t)$) denote the amount of time the tth task attempt takes when engaged in imitation learning (respectively, innovation learning); we require for both $j \in \{im, in\}$ that $\lim_{k\to\infty} \sum_{t=1}^{k} \Delta_j(t) = \infty$. Then, we assume that imitation is (weakly) faster than innovation: that $\Delta_{im}(t) \leq \Delta_{in}(t)$ for all $t \in \mathbb{N}$. Moreover, we denote by

$$T_j(i) = \sum_{n=1}^{i} \Delta_j(n) \tag{16}$$

the total amount of time that a task of learning type j occupies until the end of the ith attempt.

With sufficient time in a fixed environment, natural selection is likely to maximize the objective function (fitness) within the space of feasible policies (fitness landscape). A *policy* is defined by a function $\pi : \mathcal{H} \to \mathcal{A}$, where $\mathcal{A}$ denotes the space of feasible actions;

$$\mathcal{H} = \{(O_1, A_1, \ldots, O_{T-1}, A_{T-1}, O_T) : O_i \in \mathcal{O}, A_i \in \mathcal{A}, \text{ and the history is feasible}\}, \tag{17}$$

the space of feasible histories; $\mathcal{O}$, the space of feasible observations; and a history

$$h = (O_1, A_1, \ldots, O_{T-1}, A_{T-1}, O_T) \tag{18}$$

is called feasible if its sequence of observations and actions can occur in the model. It remains to specify the student's action space $\mathcal{A}$, observation space $\mathcal{O}$, and the objective function $V(\pi)$ on the space of policies π.

The student's objective function $V(\pi)$ is the expectation of the total payoff. Most of it comes from the payoffs yielded by the student's task attempts. Suppose that the student finishes a task attempt of time length Δ while at level of knowledge b for a task of difficulty value a. At time T that ends a learning period, the student obtains an expected payoff proportional to $f(a, b)$, scaling with the length Δ of the learning period, and simultaneously accounting for exponential time-discounting. The marginal payoff is obtained as a high-variance probabilistic lottery $\varphi(a,b) \in \mathcal{P}(S)$ with expected value $\mathbb{E}[\varphi(a,b)] = f(a,b)$. Specifically, a payoff value $\bar{s}$ is drawn independently from $\varphi(a,b)$ to determine the payoff of the task attempt

$$v(a,b,\Delta,T) = \bar{s} \int_{T-\Delta}^{T} \delta^t dt, \tag{19}$$

where $\delta \in (0,1)$ denotes the factor of exponential time-discounting. We see that the expected payoff yielded by the task attempt is

$$\mathbb{E}[v(a,b,\Delta,T)] = f(a,b) \int_{T-\Delta}^{T} \delta^t dt = \begin{cases} f(a,B(i)) \int_{T-\Delta}^{T} \delta^t dt & \text{if } b = B(i) < a, \\ \int_{T-\Delta}^{T} \delta^t dt & \text{if } b = a, \end{cases} \tag{20}$$

Instantaneously after the acquisition of this payoff at time T, the student's level of knowledge jumps to the next discrete level of knowledge $B(\cdot)$ or to the maximum level of knowledge a for the task, whichever is smaller. The expected sum of the student's task-attempt payoffs over all time $T \in [0,\infty)$ is the main component of the student's objective function $V(\pi)$.

There are three auxiliary components of the student's objective function $V(\pi)$. The first such component is as follows. After obtaining the payoff of expected value $v(a,b,\Delta,T)$, the student has the option of committing the observed payoff value to memory. Doing so requires the student to pay an expected cost of $-C_{\text{retain}}$, which represents various ecological fitness risks that result from overcommitting attention to the retention of high-variance payoff data. Due to the exponential time-discounting, the true value of the expected cost as applied to the student's objective function $V(\pi)$ is

$$-\delta^T C_{\text{retain}}, \tag{21}$$

where T denotes the ending time of the task attempt that has yielded the given payoff.

The second auxiliary component of the student's objective function $V(\pi)$ relates to a choice (described in Subsection 2.1) that the student makes before every task attempt: whether to allocate a fraction r of the task attempt's time—and the corresponding fraction of its payoff—to an alternative foraging opportunity unrelated to the task. Like in the classical Bayesian model of Subsection 2.1, the marginal payoff $s \in S$ of the alternative foraging opportunity is drawn i.i.d. from a distribution $\psi \in \mathcal{P}(S)$ and known to the student prior to their decision. If the student chooses to forgo a fraction of the task attempt's time for this alternative foraging opportunity, their payoff is changed from (19) to

$$rs \int_{T-\Delta}^{T} \delta^t dt + (1-r)v(a,b,\Delta,T) = (rs + (1-r)\bar{s}) \int_{T-\Delta}^{T} \delta^t dt. \tag{22}$$

These unrelated foraging opportunities allow the student to increase their expected payoff $V(\pi)$ strictly above the baseline level provided by the sum of the task-attempt payoffs $v(a,b,\Delta,T)$. Consequently, the student is incentivized to accurately estimate each task-attempt's payoff—as best as allowed by their informational constraints—prior to deciding whether to exploit an unrelated foraging opportunity instead.

The third auxillary component of the student's objective function $V(\pi)$ relates to the student's other choice of action. In between task attempts, the student not only chooses whether to exploit an unrelated foraging opportunity before each task attempt, but also chooses whether to quit on their current task package for an alternative one. If the student chooses to cut their losses on a given foraging task and/or their role model for it, they can choose a new task package (j,a). All of the student's task packages (j,a), including the initial one and any intermediate ones assigned after quitting, are drawn i.i.d. from the probability distribution μ defined in (15).

In addition to the option of quitting the current task, the student is also assumed to situationally possess the option of paying a fitness cost to ascertain their current task package's learning method $j \in \{im, in\}$, on which they can base their specific decision. We propose that humans carry out this ascertainment via a mental experiment to measure the length of time $\Delta_j(t)$, which may be sufficient to distinguish the speeds of the two learning methods. Specifically, our assumption that $\Delta_{im}(t) \leq \Delta_{in}(t)$ can be divided into two possibilities: $\Delta_{im}(t) < \Delta_{in}(t)$ and $\Delta_{im}(t) = \Delta_{in}(t)$. In the case of the former, a time-measurement experiment can identify the learning type j. In the case of the latter, however, it cannot. Each mental time-measurement experiment requires the student to pay an expected cost $-C_{\text{identify}}$, again due to various ecological fitness costs that can result from overloading a cognitively constrained forager's decision-making. Due to the exponential time-discounting, the true value of the expected cost as applied to the student's objective function $V(\pi)$ is

$$-\delta^T C_{\text{identify}}, \tag{23}$$

where T denotes the ending time of the task attempt during which the time-measurement experiment was performed.

We have introduced all components of the student's objective function $V(\pi)$, as well as all components of the student's action space $\mathcal{A}$. Unlike the classical Bayesian model of Subsection 2.1, our model is characterized by a potential tradeoff between earlier and later payoffs. In the classical Bayesian model, each of the agent's actions was only relevant to maximizing the payoff of the corresponding task attempt, not to any future ones. Thus, the relative weights of each task attempt's payoff do not affect the agent's decision problem. In contrast, in our model, the student has two actions—quitting the current task and identifying the learning type via a time-measurement experiment—that reduces payoffs in the short-term for a potential gain in long-term payoffs. Thus, specifying the relative weights of each task attempt's payoff is essential for the prescription of the optimal policy π. As is standard, we have set these relative weights to be exponentially decaying in time, which aids model tractability and captures the evolutionary fact that earlier payoffs are likelier to be relevant to fitness than later payoffs.

Formally, the student's actions are of the form

$$A_t = (x_{\text{forgo}}(t), x_{\text{identify}}(t), x_{\text{retain}}(t), x_{\text{quit}}(t)), \tag{24}$$

where

$$x_{\text{forgo}}(t) : S \to \{\text{true}, \text{false}\} \tag{25}$$

denotes the choice of whether to forgo a fraction of the tth task attempt's time to exploiting an alternative foraging opportunity of a known marginal payoff $s \in S$;

$$x_{\text{identify}}(t) : S \to \{\text{true}, \text{false}\} \tag{26}$$

denotes the choice of whether to pay an expected cost of $-C_{\text{identify}}$ to identify the learning type $j \in \{im, in\}$ during the tth task attempt via a time-measurement experiment, given the alternating foraging opportunity's previously drawn marginal payoff s;

$$x_{\text{retain}}(t) : S \times S \to \{\text{true}, \text{false}\} \tag{27}$$

denotes the choice of whether to retain the observation of the tth task attempt's payoff given $(s, \bar{s}) \in S \times S$, where s is given as above and $\bar{s}$ denotes the task-specific marginal payoff; and

$$x_{\text{quit}}(t) \in \{\mathcal{K}(s, \bar{s}, j, c) : S \times (S \cup \{\text{null}\}) \times \{\text{im}, \text{in}, \text{null}\} \times \{\text{true}, \text{false}\} \to \{\text{true}, \text{false}\}\}$$

denotes the choice of whether to quit the current task after the tth task attempt. When the student has not performed the identification of the learning type j during the current task attempt, $x_{\text{identify}}(t) = \text{false}$, then the value $x_{\text{quit}}(t)$ takes the form of a boolean-valued function $\mathcal{K}(s, \bar{s}, \text{null}, c)$: a function of the alternative foraging opportunity's marginal payoff s; of the task's yielded marginal payoff $\bar{s}$ (which may be unretained and thus given by $\bar{s} = \text{null}$); and whether or not the level of knowledge has caught up to the task difficulty a, denoted by

$$c \in \{\text{true}, \text{false}\}. \tag{28}$$

If $c = \text{true}$, then we say that learning has completed during this task attempt. In the opposite case of $x_{\text{identify}}(t) = \text{true}$, $x_{\text{quit}}(t)$ takes the form of a boolean-valued function $\mathcal{K}(s, \bar{s}, j, c)$ for $j \in \{im, in\}$, representing the decision whether to quit conditional on the identified learning type being imitation or innovation, on the payoff observation, and on whether learning has completed during this task attempt. We also note the feasibility constraint that the value $x_{\text{identify}}(t)$ is required to satisfy the feasibility constraint that $x_{\text{identify}}(t) = \text{true}$ is only possible if $\Delta_{im}(t) < \Delta_{in}(t)$ rather than $\Delta_{im}(t) = \Delta_{in}(t)$.

The student's observations are of the form

$$O_t = (b(t), x_{\text{type}}(t), x_{\text{payoff}}(t)), \tag{29}$$

where

$$b(t) \in [0, \infty) \tag{30}$$

denotes the level of knowledge after the tth task attempt;

$$x_{\text{payoff}}(t) \in S \cup \{\text{null}\} \tag{31}$$

denotes the student's observed payoff value (if the payoff observation was not retained, then we use the denotation "null"); and

$$x_{\text{type}}(t) \in \{\text{null}, \text{im}, \text{in}\} \tag{32}$$

denotes whether the student has carried out a mental identification of the learning type during the tth task attempt (if this is false, then we use the denotation "null"), and if so, whether the result was imitation (im) or innovation (in).

In summary, Table 1 provides the list of parameters comprising our learning model, and

Table 2 presents a step-by-step algorithm for the model. The expected payoff of the policy π (correcting for time-discounting) during the time remaining after a history h is given by

$$V_h(\pi) = \mathbb{E}\Bigg[\sum_{k=0}^{\infty}\Bigg(\Big(rx_{\text{forgo}}(k)s(k) + (1 - rx_{\text{forgo}}(k))\bar{s}(k)\Big)\int_{T(k)}^{T(k+1)}\delta^t dt \\ - \delta^{T(k+1)}\left(C_{\text{retain}}x_{\text{retain}}(k) + C_{\text{identify}}x_{\text{identify}}(k)\right)\Bigg)\Bigg], \quad (33)$$

where we have abused notation by having $\bar{s}(k)$, $s(k)$, and the choices $x_{\Box}(k)$ denote the values of $\bar{s}, s,$ and the choices $x_{\cdot}$ during the kth learning period from the present, letting $T(k)$ denote the ending time of the kth learning period from the present, and setting the boolean values of the choices $x_{\Box}(k)$ to be 0 when false and 1 when true.

Given a choice of parameters, the corresponding model parametrization $\boldsymbol{M}$ can be solved numerically with dynamic-programming-type methods. However, we instead pursue an analytic study to demonstrate desired facts about the model that hold more generally, regardless of the specific choice of parameters. The results of this investigation are documented in Section 3.

3 Results

We denote the space of feasible policies of the model described in Subsection 2.2 by Π. A policy π is called optimal if it maximizes the expected payoff in the remaining time at any feasible history h:

$$\pi \in \arg\max_{\pi\in\Pi} V_h(\pi). \quad (34)$$

In the followinwg, we obtain results on properties necessarily possessed by any optimal policy π, which can help simultaneously explain the various empirically documented deviations of human confidence from a classically Bayesian estimate of past payoff data.

First, if the magnitude C_{retain} of the expected cost of retaining payoff observations is sufficiently large, then no optimal policy π ever retains payoff observations. This can be seen, for example, by taking

$$C_{\text{retain}} > \int_0^{\infty}\delta^t \max(S)dt, \quad (35)$$

an upper bound—for any time T at which a task attempt ends—to the payoff (accounting for time-discounting) that can be obtained during the remaining time. The upper bound (35) is obtained when the student receives the maximal marginal payoff $\max(S)$ for every task, and does not pay any cost to retaining payoff observations or identifying the task's learning type. If C_{retain} were larger than this maximum possible expected payoff in the remaining time, then the information yielded by paying a cost of that magnitude would clearly never be worth it.

Throughout this paper, we assume that the magnitude C_{retain} of the expected cost of "observing" (in the ecological setting, retaining in memory) payoff data is great enough that the student does not ever do so: so that the optimal choice $x_{\text{retain}}(t)$ is always given by

$$x_{\text{retain}}(t) = \text{false}. \tag{36}$$

This is functionally equivalent to assuming that the payoff data is unavailable to the student.

The second characteristic that an optimal policy π must possess is the following. Every action $\pi(h)$ of an optimal policy in response to a history h might as well solely depend on the information of h relevant to the current task (j, a), and not on the other information (relevant to the previous tasks); this follows from the assumption that the student's tasks are statistically independent. Specifically, the choices of $x_{\text{forgo}}(t), x_{\text{identify}}(t)$, and $x_{\text{quit}}(t)$ should only depend on the conditional distribution $\mu_{\text{cond}}(h)$ of the current task's value (j, a), conditional on the information contained in the past history h. This information, which allows the student to rule out (via Bayes' formula of conditional probability) certain task values (j, a) from the initial conditional distribution of μ, includes two components. For one thing, if there has been a time-measurement experiment on the current task, say with result $j \in \{im, in\}$, then the student can rule out all task values (j', a) with $j' \neq j$.

For another, the student's past sequence of knowledge levels on the task, $b(0), b(1), \ldots, b(i-1)$, allows the student to rule out task values. If the sequence ends in one or more instances of $b = a \notin \{B(i) : i \in \mathbb{N}\}$, then the student knows that their level of knowledge b has caught up to the maximum value a. In other words, all task values (j', a') with $a' \neq a$ can be ruled out. However, if the sequence has been completely consistent with the discrete knowledge values $\{B(i) : i \in \mathbb{N}\}$ of the model, then the only task values (j', a') that can be ruled out are those with $a' \leq b = B(i)$. (Without loss of generality, we assume that the probability-zero event that the task difficulty a drawn from μ precisely equals one of the model's discrete knowledge levels $B(i)$, rather than falling between them, does not occur.)

Third, in an optimal policy π, every decision $x_{\text{forgo}}(t)$ whether to forgo a fraction of a task attempt's time for a known marginal payoff of $s \in S$ must be of the form described in Subsection 2.1: forgo if the task attempt's expected marginal payoff

$$\mathbb{E}_{(j,a)\rightsquigarrow\mu_{\text{cond}}(h)}\left[f(a,b)\right] \tag{37}$$

is greater than the alternative marginal payoff s, and do not forgo if the latter is greater than the former (when they are equal, both choices are optimal). In other words, the student should choose the payoff that is greater in expectation. We call the quantity (37) the *expected marginal payoff function* or the *confidence function*. We propose that the evolutionary pressure to optimally exploit alternative foraging opportunities shaped ancestral humans' task-specific notion of confidence to track the task's expected marginal payoff (37), conditional on both the information known so far and the parameters of the ancestral environment.

The task's expected marginal payoff (37) is a function of the student's two relevant pieces of information: their level of knowledge b and their information set on the learning type j (whether they have ruled out the event $\{j = im\}$, the event $\{j = in\}$, or neither). Specifically, the confidence function can be written as $\hat{g}(E_b, E_j)$ mapping the domain

$$\begin{aligned}&(\{\{b = B(i)\} : i \in \mathbb{N}\} \cup \{b = a \neq B(i)\})\\ &\quad \times \{\{j = im \text{ ruled out}\}, \{j = in \text{ ruled out}\}, \{\text{neither } j \text{ ruled out}\}\} \ni (E_b, E_j)\end{aligned}$$

to the range of marginal payoffs $[0, 1]$, where E_b denotes the information set regarding the student's information on b and E_j, the information set regarding the student's information on j. We compute that the confidence function (37) is generally given by

$$\hat{g}(E_b, E_j) = \begin{cases} 1 & \text{if } E_b = \{b = a \neq B(i)\}, \\ g_{im}(B(i)) & \text{if } E_b = \{b = B(i) < a\} \text{ and } E_j = \{j = in \text{ ruled out}\}, \\ g_{in}(B(i)) & \text{if } E_b = \{b = B(i) < a\} \text{ and } E_j = \{j = im \text{ ruled out}\}, \\ g_u(B(i)) & \text{if } E_b = \{b = B(i) < a\} \text{ and } E_j = \{\text{neither } j \text{ ruled out}\}, \end{cases} \tag{38}$$

for

$$g_{im}(b) = \frac{\int_{a>b} f(a,b)d\mu_{im}(a)}{\int_{a>b} d\mu_{im}(a)}, \tag{39}$$

$$g_{in}(b) = \frac{\int_{a>b} f(a,b)d\mu_{in}(a)}{\int_{a>b} d\mu_{in}(a)}, \tag{40}$$

and

$$g_u(b) = \frac{\int_{a>b} f(a,b) d\bar{\mu}(a)}{\int_{a>b} d\bar{\mu}(a)}, \tag{41}$$

where $\bar{\mu}$ denotes the probability distribution $P \circ \mu : [0, \infty) \to [0, 1]$ for the projection map $P(j, a) = a$. We call g_{im}, g_{in}, and g_u the *imitation-learning confidence function*, the *innovation-learning confidence function*, and the *unconditional confidence function*, respectively.

Let ρ_y be a distribution of the form $\rho_y(a) = (1 - y)\eta^a \log \frac{1}{\eta}$ for finite a and $\rho_y(\infty) = y$, where $y \in [0, 1)$. Define the *generalized confidence function* $g_{\rho_y} : [0, \infty) \to [0, 1]$ by

$$g_{\rho_y}(b) = \frac{\int_{a>b} f(a,b) d\rho_y(a)}{\int_{a>b} d\rho_y(a)}. \tag{42}$$

Then, we see that

$$\mu_{im} = \rho_0, \ \mu_{in} = \rho_p, \text{ and } \ \bar{\mu} = \rho_{(1-q)p}, \tag{43}$$

and therefore,

$$g_{im}(b) = g_{\rho_0}(b), \ g_{in}(b) = g_{\rho_p}(b), \text{ and } \ g_u(b) = g_{\rho_{(1-q)p}}(b). \tag{44}$$

One can then verify the following fact.

Proposition 1. *For any $b > 0$, the value of the generalized confidence function, $g_{\rho_y}(b)$, is strictly monotonically decreasing in y. In particular, the innovation-learning confidence function $g_{in}(b)$ is at most the unconditional confidence function $g_u(b)$, which is at most the imitation-learning confidence function $g_{im}(b)$. Specifically, we have*

$$g_{in}(b) \leq g_u(b) \leq g_{im}(b), \tag{45}$$

where the first inequality occurs with equality if and only if $q = 0$ (or $p = 0$, if this is allowed); and the second inequality, if and only if $q = 1$ (or $p = 0$, if this is allowed).

In other words, the evolutionarily optimal estimate of confidence at a level of knowledge b (conditional on learning not yet having completed) is decreasing in the proportion y of unlearnable tasks. This is due to the fact that the risk of unlearnability, of the task difficulty $a = \infty$, has a reduction effect on the expected marginal payoff. This risk occurs with the highest probability within the distribution of task difficulties $a > b$ conditional on $j = in$, occurs with zero probability within the distribution conditional on $j = im$, and occurs with an in-between probability value within the distribution that is unconditional of the learning type j. Thus, the reduction effect on the confidence function also falls in this order. This phenomenon is illustrated in the plots of the three confidence functions for several model parametrizations in Figure 1.

Another consequence of the risk of unlearnability is non-monotonicity. Specifically, we will show that $g_{im}(b)$ is monotonically increasing in b under a non-restrictive assumption on the marginal payoff function $f(a, b)$. Note that if all tasks were learned by imitation rather than innovation ($q = 1$), then the confidence function (37) is of the form

$$\hat{g}(E_b, E_j) = \begin{cases} 1 & \text{if } E_b = \{b = a \neq B(i)\}, \\ g_{im}(B(i)) & \text{if } E_b = \{b = B(i)\}. \end{cases} \tag{46}$$

and consequently, monotonically increasing in the level of experience i. In other words, if human confidence evolved in an environment where all tasks were learned by imitation, then we should expect it to be monotonically increasing in the level of knowledge: and thereby, the level of experience. The empirically documented confidence is non-monotonic in the level of experience, and thus unlikely to have evolved in such an environment.

On the other hand, we will show that due to the nontrivial risk of unlearnability, the confidence functions $g_{in}(b)$ and $g_u(b)$ each decay to zero as $b \to \infty$. This opens up the possibility for the confidence function (37) to be non-monotonic in the empirically documented way: general increase with an intermediate period of decrease with respect to the level of experience. Whether this non-monotonicity evolves depends on the two remaining actions prescribed by the student's optimal policy π: identifying the learning type, $x_{\text{identify}}(t)$; and quitting, $x_{\text{quit}}(t)$.

3.1 Imitation learning alone cannot explain non-monotonic confidence

Under reasonable assumptions on the model parameters, whether each of the confidence functions $g_{im}(b), g_{in}(b)$, and $g_u(b)$ is monotonic is determined by the presence of the risk of unlearnable tasks. Since the distribution μ_{im} has zero probability on the event $\{a = \infty\}$, its associated confidence function $g_{im}(b)$ is monotonically increasing in b, as long as the payoff function $f(a, b)$ satisfies the following condition:

Assumption 1. *For all $m > 0$ and $a \geq m$, the payoff function $f(a, b)$ satisfies*

$$\frac{\partial}{\partial a} f(a, a - m) > 0. \tag{47}$$

We argue that Assumption 1 is plausible because a fixed amount m of knowledge constitutes a larger fraction of the total knowledge of an easy task than a difficult task; consequently, the argument goes, not knowing it should cause a harsher penalty in the former case. However, whether this claim generally holds is a question that should be studied empirically. Note that the assumption is satisfied by the example family of payoff functions (10), but not by the example family of payoff functions (12). Our aforementioned argument would then suggest that the former family (polynomial growth) is plausible as the marginal payoff function of ancestral learning environments, but not the latter family (exponential growth).

On the other hand, since the distributions μ_{in} and $\bar{\mu}$ have a positive probability on the event $\{a = \infty\}$, their associated confidence functions $g_{in}(b)$ and $g_u(b)$ are non-monotonic. Specifically, both $g_{in}(b)$ and $g_u(b)$ decay to zero for all sufficiently large b. In fact, the functions are strictly decreasing to zero for all sufficiently large b, as long as the following condition holds.

Assumption 2. *As $b \to \infty$, the payoff function $f(a, b)$ satisfies*

$$\int_{a>b} \frac{\partial}{\partial b} f(a, b)\eta^a da \ll \eta^b. \tag{48}$$

Here, the notation $F(b) \ll G(b)$ denotes the asymptotic condition that $F(b)/G(b) \to 0$ as the input variable $b \to \infty$. Note that Assumption 2 is satisfied by the family of payoff functions (10) for any parameter $\eta \in (0, 1)$.

We summarize the above discussion in the following theorem statement.

Proposition 2. *The generalized confidence function g_{ρ_y} satisfies the following:*

a) If $y = 0$, then we have $\frac{d}{db} g_{\rho_y}(b) > 0$ for all $b \geq 0$, as long as Assumption 1 holds.

b) If $0 < y < 1$, then we unconditionally have $g_{\rho_y}(b) \to 0$ as $b \to \infty$.

c) If $0 < y < 1$, then we have $\frac{d}{db} g_{\rho_y}(b) < 0$ for all sufficiently large b, as long as Assumption 2 holds.

The expected marginal payoff of a task is monotonically increasing when there is no risk that the task is unlearnable, as is the case when it is learned by innovation. In other words, since $\mu_{im} = \rho_0$, the function g_{im} should be monotonically increasing. However, there is a nontrivial probability y of unlearnability when the learning type of the task is either uncertain or fully determined as innovation: $\mu_{in} = \rho_p$ and $\bar{\mu} = \rho_{(1-q)p}$. In this case, the corresponding expected marginal payoffs (g_{in} and g_u, respectively) both eventually monotonically decrease to zero.

We have plotted in Figure 1 confidence functions of an example model parametrization with varying marginal payoff function $f(a, b)$, which illustrate the conclusions of Proposition 2. We note in particular that functions of the form $f(a, b) = (b/a)^{\lambda}$—detailed in (10)—satisfy Assumption 1. Thus, by Proposition 2(a), any model parametrization with this choice of marginal payoff function will have a strictly increasing imitation-learning confidence function $g_{im}(b)$. However, functions of the form $f(a, b) = \zeta^{a-b}$—detailed in (12)—do not satisfy Assumption 1, which opens up the possibility that $g_{im}(b)$ will not be strictly increasing. In fact, we then can apply the change of variables $\bar{a} = a - b$ to see that the imitation-learning confidence function

$$\begin{aligned} g_{im}(b) = \frac{\left(\log \frac{1}{\eta}\right) \int_{a>b} \zeta^{a-b} \eta^{a} da}{\left(\log \frac{1}{\eta}\right) \int_{a>b} \eta^{a} da} &= \left(\log \frac{1}{\eta}\right) \int_{a>b} \zeta^{a-b} \eta^{a-b} da \\ &= \left(\log \frac{1}{\eta}\right) \int_{0}^{\infty} \zeta^{\bar{a}} \eta^{\bar{a}} d\bar{a} \end{aligned} \tag{49}$$

is constant with respect to b. Thus, we see that Assumption 1 constitutes a nontrivial necessary condition for $g_{im}(b)$ to be strictly increasing.

3.2 Analyzing a subfamily of model parametrizations via approximation

We have solved for the optimal choice of $x_{\text{forgo}}(t)$, the decision of when to forgo a proportion of the task payoff for an alternative foraging opportunity. Assuming the policy π always uses this optimal choice, the only other components of π that can vary are $x_{\text{identify}}(t)$, the decision whether to perform a time-measurement experiment to identify the learning type j; and $x_{\text{quit}}(t)$, the decision whether to quit. Recall that the only information that is relevant for the optimal choice of these components is the pair of information sets E_b and E_j regarding the the student's current task. We abuse notation by letting

$$\pi(E_b, E_j) = (x_{\text{identify}}, x_{\text{quit}}) \tag{50}$$

denote the action of the optimal policy π (omitting the components x_{retain} and x_{forgo}, which have already been solved previously) at the pair of information sets (E_b, E_j).

We proceed to define a tractable subfamily of parametrizations of our model for which the optimal estimate of confidence, as a function of the level of experience i, displays the empirically documented non-monotonicity: general increase with an intermediate period of decrease.

Whether this non-monotonicity occurs would depend, in general, on the action components $x_{\text{retain}}(k)$ and $x_{\text{forgo}}(k)$ of the optimal policy π. Our subfamily of model parametrizations will be constructed—via approximation—to have the appropriate optimal action components x_{retain} and x_{forgo} that guarantee the desired non-monotonicity.

Let us fix all choices of model parameters with the exception of the discrete knowledge jumps $\{B(i,n) : i \in \mathbb{N}, i > 0\}$, the learning period lengths $\{\Delta_j(i,n) : i \in \mathbb{N}, i > 0\}$—and the corresponding cumulative learning period lengths $\{T_j(i,n) : i \in \mathbb{N}, i > 0\}$—for $j \in \{im, in\}$, the fraction of time $r(n)$ of task attempts that can be devoted to alternative foraging opportunities, and the expected cost $C_{\text{identify}}(n)$ of a time-measurement experiment to identify the learning type j. This gives a sequence of model parameterizations $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ varying with n. We will construct $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ so that as $n \to \infty$, the *imitation-learning knowledge function* and the *innovation-learning knowledge function*, defined respectively by

$$L_{im,n}(t) = B\left(\max\{i : T_{im}(i,n) \leq t\}\right) \tag{51}$$

and

$$L_{in,n}(t) = B\left(\max\{i : T_{in}(i,n) \leq t\}\right), \tag{52}$$

can be well-approximated by the *continuous imitation-learning knowledge function*

$$L_{im,\infty}(t) : [0,\infty) \to [0,\infty), \tag{53}$$

and the *continuous innovation-learning knowledge function*

$$L_{in,\infty}(t) : [0,\infty) \to [0,\infty), \tag{54}$$

respectively. The knowledge functions $L_{im,\infty}(t)$ and $L_{in,\infty}(t)$ are required to be bijective, continuous, and piecewise continuously differentiable such that their respective derivatives $\frac{d}{dt}L_{im,\infty}(t)$ and $\frac{d}{dt}L_{in,\infty}(t)$ are positive whenever they are well-defined. We will describe the context of this continuous approximation in Subsection 3.4.

We now formally define the *continuous learning model*, a continuous approximation of our discrete learning model defined in Subsection 2.2. Suppose that instead of obtaining discrete payoffs at the end of discrete task attempts, the student obtains a flow payoff

$$\delta^t f(a(t), b(t))dt, \tag{55}$$

based on the task difficulty a and the student's level of knowledge b. The term $a(t)$ denotes the difficulty level of the task that is being learned at time t, and thus has zero derivative everywhere except for the discrete set of points of time at which tasks are quit. When a task is quit at time t, and at the starting time $t = 0$, a task is drawn i.i.d. from the distribution μ as in the model of Subsection 2.2; and if $t > 0$, the term $a(t)$ is updated to the newly drawn task difficulty.

The term $b(t)$ denotes the student's level of knowledge, and in the continuous learning model, updates continuously in the amount of time t. Specifically, we have

$$b(t) = \begin{cases} L_{im,\infty}(\bar{t}) & \text{if } j = im \text{ and } L_{im,\infty}(\bar{t}) < a \\ a & \text{if } j = im \text{ and } L_{im,\infty}(\bar{t}) \geq a \\ L_{in,\infty}(\bar{t}) & \text{if } j = in \text{ and } L_{in,\infty}(\bar{t}) < a \\ a & \text{if } j = in \text{ and } L_{in,\infty}(\bar{t}) \geq a, \end{cases} \tag{56}$$

where

$$\bar{t} = t - T_{\text{start}}(t) \tag{57}$$

denotes the length of the time period $[T_{\text{start}}(t), t]$ spent learning the current task (at time t) and

$$T_{\text{start}}(t) \tag{58}$$

denotes the time at which the current task has been drawn.

We further suppose that in the continuous learning model, there is no option to exploit alternative foraging opportunities. Similarly, we suppose that the learning type of a task is not information that can be learned by paying a cost. The justification for these assumptions is that these quantities—the payoff difference due to alternative foraging opportunities and the costs of identifying the learning type—become negligible as $n \to \infty$ in the continuous approximation.

Finally, we suppose that the option to quit for an opportunity-cost task satisfies the following. For a positive constant β, the student can—when learning has not yet completed—either quit all tasks (both $j = im$ and $j = in$) at any level of experience $b \in (0, \infty)$ without identifying the task type, or quit $j = im$ tasks at a level of experience $b_{im} \in [\beta, \infty) \cup \{\infty\}$ and $j = in$ tasks at a level of experience $b_{in} \geq [\beta, \infty) \cup \{\infty\}$.

The student's strategy space in the continuous learning model pertains entirely to quitting, and is given by

$$\mathcal{A}_\infty = \mathcal{Q}^\infty = \left(((0, \infty) \cup \infty) \cup ([\beta, \infty) \cup \{\infty\})^2\right)^\infty \tag{59}$$

for

$$\mathcal{Q} = ((0,\infty) \cup \infty) \cup ([\beta,\infty) \cup \{\infty\})^2 . \tag{60}$$

Here, the first subset $(0,\infty)$ denotes the set of quitting strategies b that quit all tasks at any level of experience $b > 0$ without identifying the learning type, and the second subset $([\beta,\infty) \cup \{\infty\})^2$ denotes the set of quitting strategies (b_{im}, b_{in}) that quit $j = im$ tasks at a level of experience $b_{im} \in [\beta,\infty) \cup \{\infty\}$ and $j = in$ tasks at a level of experience $b_{in} \geq [\beta,\infty) \cup \{\infty\}$. The action

$$(\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots) \in \mathcal{A}_\infty \tag{61}$$

denotes the overall strategy that quits the ith task using the strategy action $\boldsymbol{b}_i$ for $i \in \mathbb{N}$. The total payoff in the continuous learning model is given by

$$\int_0^\infty \delta^t f(a(t), b(t))dt, \tag{62}$$

where $a(t)$ is the difficulty value of the task being learned at time t (which discretely changes whenever a new task is drawn), and $b(t)$ is the student's level of knowledge of this task.

In summary, Table 3 provides the list of parameters comprising our continuous learning model, and Table 4 provides a step-by-step algorithm for the model. The student's objective is to maximize the expected payoff, the expected value of (62):

$$V_\infty((\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots)) = \mathbb{E}\left[\int_0^\infty \delta^t f(a(t), b(t))dt\right]. \tag{63}$$

Decision theory yields that the maximal expected payoff $V_\infty\left((\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots)\right)$ is obtained by a strategy that acts in the same way for every history sharing the same information set. In particular, the maximal payoff is obtained by a strategy that uses the same quitting strategy $\boldsymbol{b} \in \mathcal{Q}$ for every drawn task, corresponding to the strategy

$$(\boldsymbol{b}, \boldsymbol{b}, \ldots) \in A_\infty. \tag{64}$$

The expected total payoff of such a quitting strategy $\boldsymbol{b}$ is given by the function

$$V_\infty(\boldsymbol{b}) = \begin{cases} V_{\infty,u}(b) & \text{if } \boldsymbol{b} = b, \\ V_{\infty,c}(b_{im}, b_{in}) & \text{if } \boldsymbol{b} = (b_{im}, b_{in}). \end{cases} \tag{65}$$

Here, the value function $V_{\infty,u}(b)$ is defined by

$$V_{\infty,u}(b) = qV_{im,\infty,u}(b) + (1-q)V_{in,\infty,u}(b), \tag{66}$$

where $(V_{im,\infty,u}(b), V_{in,\infty,u}(b))$ is the solution to the system of equations

$$\begin{aligned} V_{im} =& \int_0^b \left(\int_0^{L_{im,\infty}^{-1}(a)} \delta^t f(a, L_{im,\infty}(t))dt + \int_{L_{im,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{im}(a) \\ &+ \int_{a>b} \left(\int_0^{L_{im,\infty}^{-1}(b)} \delta^t f(a, L_{im,\infty}(t))dt + \delta^{L_{im,\infty}^{-1}(b)}(qV_{im} + (1-q)V_{in}) \right) d\mu_{im}(a), \end{aligned} \tag{67}$$

and

$$\begin{aligned} V_{in} =& \int_0^b \left(\int_0^{L_{in,\infty}^{-1}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{in}(a) \\ &+ \int_{a>b} \left(\int_0^{L_{in,\infty}^{-1}(b)} \delta^t f(a, L_{in,\infty}(t))dt + \delta^{L_{in,\infty}^{-1}(b)}(qV_{im} + (1-q)V_{in}) \right) d\mu_{in}(a); \end{aligned} \tag{68}$$

while the value function $V_{\infty,c}(b_{im}, b_{in})$ is defined by

$$V_{\infty,c}(b_{im}, b_{in}) = qV_{im,\infty,c}(b_{im}, b_{in}) + (1-q)V_{in,\infty,c}(b_{im}, b_{in}), \tag{69}$$

where $(V_{im,\infty,c}(b_{im}, b_{in}), V_{in,\infty,c}(b_{im}, b_{in}))$ is the solution to the system of equations

$$\begin{aligned} V_{im} =& \int_0^{b_{im}} \left(\int_0^{L_{im,\infty}^{-1}(a)} \delta^t f(a, L_{im,\infty}(t))dt + \int_{L_{im,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{im}(a) \\ &+ \int_{a>b_{im}} \left(\int_0^{L_{im,\infty}^{-1}(b_{im})} \delta^t f(a, L_{im,\infty}(t))dt + \delta^{L_{im,\infty}^{-1}(b_{im})}(qV_{im} + (1-q)V_{in}) \right) d\mu_{im}(a), \end{aligned} \tag{70}$$

and

$$V_{in} = \int_0^{b_{in}} \left(\int_0^{L^{-1}_{in,\infty}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L^{-1}_{in,\infty}(a)}^{\infty} \delta^t dt \right) d\mu_{in}(a) \\ + \int_{a > b_{in}} \left(\int_0^{L^{-1}_{in,\infty}(b_{in})} \delta^t f(a, L_{in,\infty}(t))dt + \delta^{L^{-1}_{in,\infty}(b_{in})}(qV_{im} + (1-q)V_{in}) \right) d\mu_{in}(a). \tag{71}$$

In fact, we can explicitly solve for these value functions.

Lemma 3. *The value functions $V_{im,\infty,c}$, $V_{in,\infty,c}$, $V_{im,\infty,u}$, and $V_{in,\infty,u}$ are given by*

$$((V_{im,\infty,c}(b_{im}, b_{in}), V_{in,\infty,c}(b_{im}, b_{in})) = \left(\hat{V}_{im}(b_{im}, b_{in}), \hat{V}_{in}(b_{im}, b_{in}) \right) \tag{72}$$

and

$$((V_{im,\infty,u}(b), V_{in,\infty,u}(b)) = \left(\hat{V}_{im}(b, b), \hat{V}_{in}(b, b) \right). \tag{73}$$

Here, the functions $\hat{V}_{im}, \hat{V}_{in} : ((0,\infty) \cup \{\infty\})^2 \to [0, \infty)$ are defined by

$$\hat{V}_{im}(b_{im}, b_{in}) = \frac{\mathfrak{d}\mathfrak{e} - \mathfrak{b}\mathfrak{f}}{\mathfrak{g}} \tag{74}$$

and

$$\hat{V}_{in}(b_{im}, b_{in}) = \frac{\mathfrak{a}\mathfrak{f} - \mathfrak{c}\mathfrak{e}}{\mathfrak{g}} \tag{75}$$

for

$$\mathfrak{a} = 1 - q\delta^{L^{-1}_{im,\infty}(b_{im})}\eta^{b_{im}} \tag{76}$$

$$\mathfrak{b} = -(1-q)\delta^{L^{-1}_{im,\infty}(b_{im})}\eta^{b_{im}}, \tag{77}$$

$$\mathfrak{c} = -q\delta^{L^{-1}_{in,\infty}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right), \tag{78}$$

$$\mathfrak{d} = 1 - (1-q)\delta^{L^{-1}_{in,\infty}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right), \tag{79}$$

$$\mathfrak{e} = \int_0^{b_{im}} \left(\int_0^{L^{-1}_{im,\infty}(a)} \delta^t f(a, L_{im,\infty}(t))dt + \int_{L^{-1}_{im,\infty}(a)}^{\infty} \delta^t dt \right) d\mu_{im}(a) \\ + \int_{a > b_{im}} \left(\int_0^{L^{-1}_{im,\infty}(b_{im})} \delta^t f(a, L_{im,\infty}(t))dt \right) d\mu_{im}(a), \tag{80}$$

$$\mathfrak{f} = \int_0^{b_{in}} \left(\int_0^{L_{in,\infty}^{-1}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{in}(a)$$

$$+ \int_{a > b_{in}} \left(\int_0^{L_{in,\infty}^{-1}(b_{in})} \delta^t f(a, L_{in,\infty}(t))dt \right) d\mu_{in}(a), \tag{81}$$

and

$$\mathfrak{g} = 1 - \delta^{L_{in,\infty}^{-1}(b_{in})} \left(p + (1-p)\eta^{b_{in}}\right) + q \left(\delta^{L_{in,\infty}^{-1}(b_{in})} \left(p + (1-p)\eta^{b_{in}}\right) - \delta^{L_{im,\infty}^{-1}(b_{im})} \eta^{b_{im}} \right). \tag{82}$$

In particular, we have $V_\infty(b_{im}, b_{in}) = \hat{V}_\infty(b_{im}, b_{in})$ *and* $V_\infty(b) = \hat{V}_\infty(b)$ *for*

$$\hat{V}_\infty = q\hat{V}_{im} + (1-q)\hat{V}_{in}. \tag{83}$$

Note that it makes sense to view the space of quitting strategies $\mathcal{Q}$ as the domain

$$\bar{\mathcal{Q}} = \{(b, b) : b \in (0, \infty) \cup \{\infty\}\} \cup \{(b_{im}, b_{in}) : b_{im}, b_{in} \geq \beta\} \subset ([0, \infty) \cup \infty)^2, \tag{84}$$

representing the space of strategies that use the same quitting strategy for every task. Note that the two subsets above nontrivially intersect. This has the meaning that the strategy $\boldsymbol{b} = b \geq \beta$ that quits without identifying the learning type obtains the same payoff as the strategy $\boldsymbol{b} = (b, b)$ that identifies the learning type before quitting, due to our assumption that the cost of identifying the learning type limits to zero in the continuous approximation.

We formalize the aforementioned assumptions regarding the approximation of the discrete learning models $\boldsymbol{M}(n)$ by the continuous learning model $\boldsymbol{M}(\infty)$. A sequence of model parametrizations $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ is said to *converge to* the continuous model parametrization $\boldsymbol{M}(\infty)$ if:

1. The sequence of functions $\{L_{j,n}\}_{n>0}$ monotonically converges (increasing with respect to n) to $L_{j,\infty}$ in a way such that $L_{j,\infty}(T(i, n)) = B(i, n)$ for all n and i.

2. The parameters $\delta, f(a, b), p, q$, and η are shared by all $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ and $\boldsymbol{M}(\infty)$.

3. We have $\Delta_{im}(i, n) = \Delta_{in}(i, n)$ for all i such that $B(i, n) < \beta$, and $\Delta_{im}(i, n) < \Delta_{in}(i, n)$ for all i such that $B(i, n) \geq \beta$.

4. The parameters $r(n)$ and $C_{\text{identify}}(n)$ are monotonically decreasing to zero such that

$$r(n) \ll C_{\text{identify}}(n) \ll 1. \tag{85}$$

The first condition constitutes the assumption that the student's accumulation of knowledge is sufficiently fine, and thus can be approximated by a continuous knowledge function. The second condition specifies the shared parameters between the approximated model parametrizations and the approximating continuous learning model. The third condition constitutes the assumption that the speeds of imitation and innovation are too similar to distinguish in the early stages of learning ($b < \beta$), but branch off so that they become distinguishable in the later stages ($b \geq \beta$). This branch-off can occur, for example, if the respective speeds of learning increase over time—as they did in the experimental variant of Sanchez and Dunning (2020) that measured learning speeds—such that the rate of increase is faster for imitation than it is for innovation. And finally, the fourth condition represents the assumption that the additional payoffs from alternative foraging opportunities are negligible compared to the ecological fitness cost of identifying a given task's learning type, which is negligible compared to task payoffs.

This notion of convergence is key to our approach of continuous approximation. Recall that the optimal payoff of our original discrete learning model is achieved by a policy π whose choice of action $\pi(h)$ is the same for all histories of the same pair of information sets (E_b, E_j). For such a policy π, define

$$i_{\text{identify}} = \min\{i : \pi(\{b = B(i)\}, \{\text{neither } j \text{ ruled out yet}\}) = (\text{true}, x_{\text{quit}})\}, \tag{86}$$

the level of experience at which the learning type j is identified. If the policy π (conditional on learning not having completed) quits earlier than i_{identify}, say at level of experience

$$i_{\text{quit,u}} = \min\{i < i_{\text{identify}} : \pi(\{b = B(i)\}, \{\text{neither } j \text{ ruled out yet}\}) = (\text{false}, \text{true})\}, \tag{87}$$

then we say that the quitting strategy of π is *representable by* $\boldsymbol{b} = B(i_{\text{quit,u}})$. If the policy π (conditional on learning not having completed) quits at or later than i_{identify}, then we define

$$i_{\text{quit,im}} = \min\{i \geq i_{\text{identify}} : \pi(\{b = B(i)\}, \{j = in \text{ ruled out}\}) = (\text{false}, \text{true})\}, \tag{88}$$

and

$$i_{\text{quit,in}} = \min\{i \geq i_{\text{identify}} : \pi(\{b = B(i)\}, \{j = im \text{ ruled out}\}) = (\text{false}, \text{true})\}, \tag{89}$$

which denote the earliest levels of experience $i \geq i_{\text{identify}}$ at which tasks of learning type j are quit (conditional on learning not having completed). Then, we say that the quitting strategy of π is *representable by* $\boldsymbol{b} = (B(i_{\text{quit,im}}), B(i_{\text{quit,in}}))$.

Assuming these conditions hold, we have the following approximation result:

Proposition 4. *Suppose we have a sequence of model parametrizations* $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ *that converges to the continuous learning model* $\boldsymbol{M}(\infty)$*. Let* V_n *denote the payoff function corresponding to* $\boldsymbol{M}(n)$*. For every* $\varepsilon > 0$*, there exists* N *sufficiently large that for all* $n \geq N$*, we have*

$$|V_n(\pi) - V_\infty(\boldsymbol{b}(\pi))| < \varepsilon \tag{90}$$

whenever π *is representable as* $\boldsymbol{b}(\pi)$*.*

The intuition is that since the magnitude of the cost of identifying the learning type C_{identify} is negligible compare to the main term, and the additional payoffs from alternative foraging opportunities are even more negligible, the main term of the payoff $V_n(\pi)$—comprised of payoffs obtained from the task—will asymptotically dominate. In the proof of Proposition 4, we will construct a function $\hat{V}_n(b_{im}, b_{in})$ that can represent this main term. A key step in the proof that the inequality (90) holds will be that the constructed function with b placed in both inputs, $\hat{V}_n(b, b)$, is continuous at $b = 0$, and that the same holds for $\hat{V}_\infty(b, b)$. This allows us to apply Dini's theorem that for a sequence of continuous functions on a compact space that monotonically converges to another continuous function on the compact space, the convergence is uniform. Dini's theorem, a tool we will use several times in this paper, is the reason we have defined the notion of convergence of model parametrizations $\boldsymbol{M}(n)$ in terms of monotonic convergence of the knowledge functions $L_{j,n}$.

Through Proposition 4, we have essentially reduced the problem of studying the action components x_{identify} and x_{quit} in sufficiently fine model parametrizations $\boldsymbol{M}(n)$ to looking at the analogous problem in the continuous approximation $\boldsymbol{M}(\infty)$. We proceed to analyze the latter in the following subsections to gain an insight on the optimal choice of whether to quit the status-quo task in a sufficiently fine model parametrization $\boldsymbol{M}(n)$, i.e., with n sufficiently large. The advantage of studying the continuous learning model $\boldsymbol{M}(\infty)$ is that it is significantly more tractable. For it, we can obtain quite general results about the optimal quitting strategy $\boldsymbol{b}$, which can manifest in the evolutionarily optimal estimate of confidence in the approximated model parametrizations $\boldsymbol{M}(n)$.

3.3 Dichotomy of quitting strategies based on the learning type

We begin by proving that tasks that are known to be learned by imitation are never optimally quit in the continuous learning model, as long as Assumption 1 holds and the knowledge function $L_{im,\infty}(t)$ is convex. The intuition is the following. First, the optimal expected marginal payoff is increasing in the level of knowledge when the task is known to be learned by imitation, due to Assumption 1. Second, tasks learned by imitation are learned at least as fast at higher levels of knowledge, by the assumption of the convexity of $L_{im,\infty}(t)$. Finally, tasks learned by innovation in expectation yield less payoff than tasks learned by imitation. Thus, quitting at any level of knowledge $b > 0$ has three negative effects on expected payoff—reducing the expected marginal payoff, slowing down learning, and replacing the current imitation-learning task with an on-expectation inferior innovation-learning task—and is thus suboptimal.

Proposition 5. *In a continuous learning model, every $\boldsymbol{b} = (b_{im}, b_{in})$ that maximizes the payoff $V_\infty(\boldsymbol{b})$ must have $b_{im} = \infty$, as long as Assumption 1 holds and the imitation-learning knowledge function $L_{im,\infty}(t)$ is convex.*

As a result, the problem of finding the quitting strategy $\boldsymbol{b} = (b_{im}, b_{in})$ that maximizes the value function $V_{\infty,c}(b_{im}, b_{in})$ becomes a one-dimensional maximization problem

$$\max_{b_{in} \in [\beta,\infty) \cup \{\infty\}} V_{\infty,c}(\infty, b_{in}). \tag{91}$$

Note that the convexity of a knowledge function $L_{j,\infty}(t)$ constitutes the assumption that knowledge-learning is (weakly) faster in its later stages. If true, this may reflect a dynamic where potential advances in task-specific knowledge are limited by the amount of previously held knowledge, so that such advances are more likely to arise from the substantial knowledge base in the late stages of learning than from the lacking knowledge base in the early stages of learning. However, the opposite assumption of a concave knowledge function $L_{j,\infty}(t)$, the assumption that knowledge-learning is (weakly) faster in its earlier stages, is also plausible. If true, this may reflect a dynamic where there are more "low-hanging fruits" in the early stages of learning than in the late stages. Empirical studies can help quantitatively investigate aspects of knowledge accumulation as a function of time: in particular, which of the two aforementioned dynamics dominates at any given stage of learning.

Next, we prove an unconditional result: that tasks known to be either learned by innovation or of ambiguous learning type are always optimally quit at an intermediate level of knowledge.

Proposition 6. *In a continuous learning model, every $\boldsymbol{b} = (b_{im}, b_{in})$ that maximizes the value function $V_\infty(\boldsymbol{b})$ satisfies $b_{in} < \infty$. Also, every $\boldsymbol{b} = b \in (0, \infty) \cup \{\infty\}$ that maximizes the value of the function $V_\infty(\boldsymbol{b})$ satisfies $b < \infty$.*

The intuition is that these tasks, in contrast to tasks known to be learned by imitation, come with a risk of unlearnability that asymptotically dominates as the level of knowledge becomes sufficiently high. As a result, conditional on learning not yet having completed, the expected payoff from staying the course asymptotically decays to the point of being overtaken by that yielded by switching to an opportunity-cost task.

3.4 Implications for the evolutionarily optimal estimate of confidence

Consider the evolutionarily optimal estimate of confidence $\hat{g}(E_b, E_j)$, defined in (38), for a model parametrization $\boldsymbol{M}(n)$ for a sufficiently large n. Unlike in the continuous limit $\boldsymbol{M}(\infty)$, the model parametrization $\boldsymbol{M}(n)$ is characterized by alternative foraging opportunities, whose exploitation factors into the payoff function $V_n(\pi)$. Thus, the student in the model $\boldsymbol{M}(n)$ is predicted to evolve the optimal estimate of confidence $\hat{g}(E_b, E_j)$.

The possible values of confidence as a function of the level of knowledge b (conditional on learning not having completed yet, $b < a$) are $g_{im}(b), g_{in}(b)$, and $g_u(b)$. Under Assumption 1 and the assumption that the imitation-learning knowledge function $L_{im,\infty}(t)$ is convex, tasks learned by imitation are never quit. Consequently, there are two possibilities for how a payoff-maximizing strategy $\boldsymbol{b}$ in the approximating continuous learning model $\boldsymbol{M}(\infty)$ will learn tasks.

The first possibility, corresponding to the case that $\boldsymbol{b} = b'$, is that tasks are learned until a level of knowledge b' and quit if learning has not completed by then. In this case, the optimal estimate of confidence $\hat{g}(E_b, E_j)$, conditional on $b < a$, is given by $g_u(b)$ for $b < b'$, and tasks are never learned to a higher level of knowledge than b'. This conclusion seems empirically untenable for two reasons. First, there are many instances of human learning of tasks that continues on to high levels of experience and knowledge without quitting. Second, the function $g_u(b)$ has been shown in Proposition 2 to eventually decay to zero for b sufficiently high, which contradicts the empirical pattern that confidence is generally increasing in the level of experience (albeit with an intermediate period of decrease).

The second possibility, corresponding to the case that $\boldsymbol{b} = (\infty, b_{im})$ for $b_{im} \in [\beta, \infty)$, is that tasks are learned until a level of knowledge b_{im}, at which point tasks of innovation-learning type are quit if learning has not completed by then and tasks of imitation-learning type are learned to completion. Recall that we have assumed that the additional payoff obtainable from alternative foraging opportunities, which scale with r, is negligible compared to the cost of identifying the learning type $-C_{\text{identify}}$. A consequence of this assumption is that in the limit $n \to \infty$, the only possible upside of identifying the learning type is to enable differentiated choices pertaining to quitting that differ between the two learning types. Moreover, the negligibility of the cost $-C_{\text{identify}}$ in comparison to payoffs from the task necessitate that this cost is paid at the latest possible time which allows for the optimal such differentiated quitting strategy to be played: specifically, during the task attempt i_{identify} for which $b_{im} = B(i_{\text{identify}}, n)$ is payoff-maximizing among the possible quitting points $\{B(i,n)\}_{n\in\mathbb{N},n>0}$.

Because of this, when the strategy of the form $\boldsymbol{b} = (\infty, b_{im})$ is used, the optimal estimate of confidence $\hat{g}(E_b, E_j)$, conditional on $b < a$, is given by $g_u(b)$ for $b < b_{im}$ and by $g_{im}(b)$ for $b \geq b_{im}$. Since $g_u(b)$ eventually decays to zero and $g_{im}(b)$ is monotonically increasing, their piecewise combination (conditional on learning not yet having completed),

$$g(b) = \begin{cases} g_u(b) & \text{if } b \leq b_{in}, \\ g_{im}(b) & \text{if } b > b_{in}, \end{cases} \tag{92}$$

can be non-monotonic in the empirically observed way: generally increasing with an intermediate period of decrease.

In order for the evolutionarily optimal estimate of confidence $\hat{g}(E_b, E_j)$ to be empirically tenable, the payoff-maximizing strategy seems to need to be of the form $\boldsymbol{b} = (\infty, b_{im})$, and not $\boldsymbol{b} = b$. To show the plausibility of the former possibility, we construct model parameters p (the proportion of unlearnable tasks among all tasks learned by innovation) and q (the proportion of tasks learned by imitation among all tasks) for which this is true. We do this by showing that both p and q can be taken sufficiently small in our continuous learning model $\boldsymbol{M}(\infty)$ so that any strategy maximizing $V_\infty(\boldsymbol{b})$ among the subset of strategies of the form $\boldsymbol{b} = b$ quits at an arbitrarily late level of knowledge b. In particular, this can be done so that b is at least β, at which point we can appeal to Proposition 5 to see that the best strategy of the form $\boldsymbol{b} = b$ is suboptimal in the overall set of strategies $\bar{\mathcal{Q}}$. Depending on the choice of model parameters (e.g., see Figure 2), the decreasing behavior at the tail end of the component function $g_u(b)$ can be captured in the piecewise function $g(b)$, where it is followed by the monotonic increase of the component function $g_{im}(b)$. Thus, it is theoretically plausible that the evolutionarily optimal estimate of confidence conditional on learning not yet having completed, $g(b)$, is generally increasing with an intermediate period of decrease.

Corollary 7. *Suppose Assumption 1 holds and the imitation-learning knowledge function $L_{im,\infty}(t)$ is convex. In the continuous learning model, fix all parameter choices except those of p and q. For every $\gamma \geq 0$, there exist choice of p and q such that the following simultaneously hold.*

a) Any quitting strategy $\boldsymbol{b} = (\infty, b_{in})$ maximizing V_∞ must satisfy

$$b_{in} > \gamma. \tag{93}$$

b) Any quitting strategy $\boldsymbol{b} = b$ maximizing $V_\infty(b)$ (where we include the limiting strategy $\boldsymbol{b} = b \to 0$ in the domain) must satisfy

$$b > \gamma. \tag{94}$$

To prove this, we will use the following lemma, a comparative-statics result which is also of independent interest. It is comprised of two intuitive facts. First, the payoff value is decreasing in the proportion p of unlearnable tasks among those learned by innovation, which makes sense because unlearnable tasks yield the minimum possible payoff. Second, the payoff value is increasing in the proportion q of tasks learned by imitation, which makes sense because these tasks on expectation yield higher payoffs than those learned by innovation.

Lemma 8. *For any fixed $(b_{im}, b_{in}) \in \bar{\mathcal{Q}} \cup \{(0,0)\}$, the following are true.*

a) We have

$$\frac{\partial}{\partial p}\hat{V}_\infty(b_{im}, b_{in}) \leq 0, \tag{95}$$

with equality if and only if $q = 1$.

b) If Assumption 1 holds and the imitation-learning knowledge function $L_{im,\infty}(t)$ is convex, then we have

$$\frac{\partial}{\partial q}\hat{V}_\infty(\infty, b_{in}) > 0. \tag{96}$$

3.5 An example showing the plausibility of non-monotonic confidence

We conclude by constructing a family of model parametrizations $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ whose approximating continuous learning model $\boldsymbol{M}(\infty)$ can be used to show that the confidence function $g(b)$

that evolves in a sufficiently fine model parametrization $\boldsymbol{M}(n)$ can plausibly be non-monotonic in the desired way: general increase with an intermediate period of decrease. The choice of parameters for $\boldsymbol{M}(n)$ is presented in Table 5. Then, the family of model parametrizations $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ is approximable by the continuous learning model $\boldsymbol{M}(\infty)$, which has knowledge functions $L_{im,n}(t)$ and $L_{in,n}(t)$ that are determined—by the values $\Delta_j(i,n)$ and $B(i,n)$—to be

$$L_{im,\infty}(t) = \begin{cases} t & \text{if } t < 2, \\ 2(t-1) & \text{if } t \geq 2, \end{cases} \tag{97}$$

which is convex; and

$$L_{in,\infty}(t) = t. \tag{98}$$

Also, the threshold for learning-type identification is determined—by the values $\Delta_j(i,n)$—to be $\beta = 2$. Moreover, all other parameters are shared with the model parametrizations $\boldsymbol{M}(n)$. Plots relevant to the family $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ and its approximating continuous learning model $\boldsymbol{M}(\infty)$ are shown in Figure 2.

We use Mathematica 12.2's `NMaximize` function to find a local-maximizing, potentially global-maximizing quitting strategy $\boldsymbol{b} = (\infty, b_{im}) \in \bar{\mathcal{Q}}$ for $b_{im} \approx 5.32$. That the quitting strategy $\boldsymbol{b} = (\infty, b_{im})$ is local-maximizing and ostensibly global-maximizing is illustrated in Figure 2(f)'s plot of the global-maximum candidates $V_\infty(b)$ for $b < 2$ and $V_\infty(\infty, b)$ for $b \geq 2$, within the domain $0 \leq b \leq 100$. Thus, it is plausible that the quitting strategy $\boldsymbol{b} = (\infty, b_{im})$ evolves, and consequently, that $b = b_{im}$ is the cutoff point for the (limiting) piecewise-defined confidence function $g(b)$ that is optimal when using the quitting strategy $\boldsymbol{b} = (\infty, b_{im})$. As shown in Figure 2(d), this cutoff point makes the confidence function $g(b)$ is non-monotonic in the desired way: general increase with an intermediate period of decrease. By Proposition 4, this type of non-monotonic pattern will manifest in the corresponding confidence functions $g(b)$ of the model parametrizations $\boldsymbol{M}(n)$ for sufficiently large n, thereby illustrating via example the theoretical plausibility of this pattern's evolution.

4 Discussion

Classical Bayesian models are often used to represent task-learning over repeated attempts, each of which yields an observable payoff (e.g., Savage, 1972). In this paper, we have described a practical test for rejecting the null hypothesis that a learner is meaningfully learning from their environmental feedback in the sense of classical Bayesian updating. The test—essentially a corollary of standard Bayesian statistics—is to check whether the learner's estimate of their expected payoff-acquisition ability is converging to the mean of the past payoff data.

However, there is extensive empirical evidence of people's persistent failures to meaningfully learn from high-variance environmental feedback. This manifests in cognitive biases like underinference, the hard-easy effect, and recurrently non-monotonic confidence. Our test thus suggests that we should consider rejecting the null hypothesis that humans by default meaningfully learn (in the sense of classical Bayesian updating) from high-variance payoff data. Indeed, the version of the classical Bayesian model we have presented in Subsection 2.1 is specialized to repeated task-learning and incorporates the realistic assumption that a cognitive biological agent bins observations into finitely many bins. Under this assumption, tasks that yield low-variance payoff data are easily learned via deterministic causal inference, because it is likely that nearly all payoff data will fall in a single observational bin. However, learning tasks that yield high-variance payoff data requires a large number of observations for classical Bayesian inference to reliably learn the true state. Overcommitting attention to meaningfully retain a large number of high-variance observations could result in onerous ecological fitness costs, which we hypothesize is the causal mechanism behind the proposed non-selection of classically Bayesian learning strategies in settings of high-variance payoff data.

Next, we have modified the classical Bayesian model to represent ancestral humans' learning environment in a way that can evolutionary explain the puzzling predictive inadequacies of classical Bayesian updating models (when applied to humans). When the ecological fitness cost of retaining payoff data is high, the optimal strategy does not retain them, in contrast to the Bayesian principle that free information should always be taken. The optimal strategy then instead relies on setting-specific sources of information, as theorized by the ecological rationality hypothesis. The informational setting of ancestral human learning is hypothesized by cultural evolutionary theory to be one where social learning of task-specific knowledge is paramount.

Our modified Bayesian model seeks to represent this hypothesized learning environment. In it, a student attempts to learn a fitness-relevant task via attempted imitation of a role model, with the option of switching between tasks and role models (between task packages). The main term of the student's payoff function is comprised of payoffs yielded by task attempts, which are obtained in the form of high-variance probabilistic lotteries and thus unfeasible to meaningfully retain. However, the payoff function also has a secondary term comprised of the ecological fitness cost of identifying the learning type (we hypothesize that this is accomplished via a mental time-measurement experiment to distinguish learning speeds), as well as a tertiary term comprised of additional payoffs obtained by devoting a fraction of a task attempt's time to opportunistically exploiting alternative foraging opportunities instead.

Optimal exploitation of alternative foraging opportunities requires an accurate estimate of the task's expected marginal payoff conditional on the known information, which—in our hypothesized domain of high-variance, difficult-to-retain payoff data—is comprised of the task's learning type, if known (successful imitation versus *de facto* innovation); and their level of knowledge on the task. This evolutionarily optimal estimate of the expected marginal payoff—of the student's confidence at the task—is a piecewise function of their level of experience, whose piecewise cutoff point is determined by the optimal point at which tasks learned by *de facto* innovation are quit. In order for this confidence function to not be always monotonically increasing, it is necessary (as long as Assumption 1 holds) that not all attempted imitation learning is successful: that a positive proportion of tasks are learned instead via *de facto* innovation.

Moreover, we demonstrate that this confidence function can be non-monotonic in the specifically desired way: general increase with an intermediate period of decrease. This specific non-monotonic pattern, which we have demonstrated for a tractable subfamily of model parametrizations, arises because of the following interplay. Learning via *de facto* innovation while attempting to imitate a role model is not guaranteed to complete in finite time, because the task may be unlearnable. On the other hand, this risk does not exist when the student learns from authentically imitating a role model, since conditional on the imitation being authentic, the role model must have successfully learned the task beforehand. The student's optimal estimate of the task's expected marginal payoff (confidence) is monotonically increasing in the level of knowledge when it is guaranteed to be learnable in finite time, but eventually decays to zero when it may instead be impossibly difficult. We thus hypothesize that the evolutionarily optimal estimate of the expected marginal payoff can be non-monotonic due to its piecewise definition. The increasing, then decreasing portion of the expected marginal payoff function is conditional on the fact that the task may be unlearnable. The final increasing portion is conditional on having ruled out the risk of unlearnability, because the tasks to which this risk is exclusive—those learned by innovation—should optimally be quit at an intermediate level of knowledge.

In short, we hypothesize that the desired pattern of recurrent non-monotonicity evolved due to a particular interplay between the ecologically rational estimate of task-specific confidence and the ecologically rational strategy of task/role-model turnover. A necessary condition for this interplay is the dichotomy between tasks learned by imitation (for which the risk of unlearnability does not occur) and those learned by innovation (for which it does).

We emphasize that the aforementioned subfamily of model parametrizations was specifically constructed to demonstrate the theoretical plausibility of the desired non-monotonicity in an analytically tractable subset of the family of all parametrizations of our model. We anticipate that the full subset of model parametrizations whose evolutionary optimal estimate of confidence is recurrently non-monotonic in the desired way will be larger.

We are agnostic about the precise combination of adaptive and biological mechanisms by which the ecologically rational strategy (of task-payoff estimation and task/role-model turnover) in an environment of social task-specific learning was achieved. Plausible adaptive mechanisms relevant to this strategy include genetic evolution and contemporary, likely social learning. Given that people often fail to adapt their decision-making to settings of unambiguous individual learning with zero ecological fitness costs of retaining payoff data—such as those of the experiments of Sanchez and Dunning (2018, 2020)—we propose that genetic evolution plays at least a partial role in the sense of the ecological rationality hypothesis. On the other hand, cultural evolutionary theory implies that contemporary social learning may also play at least a partial role, especially given the sheer variation of relevant parameters among the myriad environments and groups humans have inhabited and moved between. The biological mechanisms through which ecologically rational strategies of social task-learning are implemented are likely neurological, but may also be partly hormonal. Future research on both the adaptive and the biological mechanisms relevant to strategies of task-payoff estimation, task/role-model turnover, and other aspects of social task-learning would potentially be fruitful.

4.1 Implications

Our model proposes to help explain in an interwoven way two related topics: the evolutionary explanation of cognitive biases, and of why people underuse high-variance environmental feedback in the selection of role models. It does so by incorporating—into the general framework of Bayesian decision theory—the cultural-evolutionary-theoretic hypothesis that the primary informational setting of ancestral human learning was the social learning of task-specific knowledge; as well as the insight of the ecological rationality hypothesis that the method by which biological cognitive agents learn from information is constrained in a setting-specific manner, such as by their ancestral environments' ecological fitness costs of overcomitting attention.

First, our model demonstrates the evolutionarily plausibility of empirically robust cognitive biases regarding confidence, and informs us of potentially useful necessary conditions and sufficient conditions for these patterns to evolve.

1. Task-specific confidence can persistently deviate from the environmental feedback, in a way that conforms to the hard-easy effect. This requires that the ecological fitness cost of retaining payoff data is nonzero, and is guaranteed to occur if the cost is sufficiently high.
2. Task-specific confidence can be recurrently non-monotonic in the desired way: general increase with an intermediate period of decrease. This requires (as long as Assumption 1 holds) that a positive proportion of attempted imitation learning is unknowingly implemented as *de facto* innovation learning, and is guaranteed to occur in our constructed subfamily of model parametrizations.

In the course of producing these desired conclusions while aiming to maintain model parsimony, our work has identified a relatively short list of environmental parameters that are potentially key to predicting certain aspects (i.e., task-specific confidence and strategies of task/role-model turnover) of descriptive human learning of a high-variance-payoff task over repeated attempts.

Also, our model augments our understanding of how role-model-selection strategies that persistently fail to meaningfully learn from certain environmental feedback evolved. Cultural evolutionary theory hypothesizes that once some capacity for cultural transmission evolved, natural selection would have favored increasingly effective strategies for cultural learning (Henrich, 2015). In this hypothesis, ancestral humans somehow achieved the threshold level of cultural-learning capacity at which cumulative cultural evolution becomes the primary selection pressure acting on cognition. After crossing this threshold, ancestral humans with a better-than-average capacity for cultural learning would have been favored by natural selection, which would then further amplify cumulative cultural evolution. Thus, gene-culture coevolution caused an autocatalytic cycle of more effective cultural-learning strategies and greater cumulative cultural evolution. A hypothesized example of such an effective cultural-learning strategy is selective social learning: the strategy of learning from preferentially chosen role models who are likely to possess better-than-average knowledge (Boyd & Richerson, 1985).

However, empirical studies have uncovered what at first appear to be surprising suboptimalities for the role-model selection strategies that humans have actually evolved. For example, students are substantially inaccurate in assessing the help provided by their teachers (Insler et al., 2021; Weinberg et al., 2009). Also, people are persistently vulnerable to maladaptive advice from role models (de Francesco, 1939; Uscinski et al., 2016; Gladwell, 2019), such as that regarding female genital cutting (Jones et al., 1999; Wagner, 2015), funerary cannibalism (Lindenbaum, 2001), unfounded shamanistic predictions (Singh, 2018), membership in an exploitative cult (Galanter, 1989), and medical pseudoscience (Scheirer, 2020). This body of evidence begs a question: why did ancestral humans evolve to not meaningfully learn from certain environmental observations relevant to the accurate assessment of role-model quality? One might presume that an informationally rational social learner would base their role-model selection on the payoff data of potential role models, and on the learner's own payoff data in the process of imitating a given role model.

Our theory contributes to explaining this phenomenon by specializing the ecological-rationality framework (in our setting, by incorporating high ecological fitness costs of retaining environmental observations) to not only the estimation of task-specific payoffs, but also the selection of tasks/role models. Specifically, in our model, these ecological fitness costs can cause role-model-selection strategies (in our model, task/role-model turnover strategies which determine when to quit the status-quo task package for a new one) based on retaining such observations to be informationally inefficient. Classically Bayesian-rational strategies, such as those of role-model selection, are much more likely to be suboptimal when environmental observations occur with high variance. Also, our model proposes explicit mechanisms by which ancestral humans—even in the absence of feasibly retainable environmental feedback—could still have plausibly evolved on-average selective role-model-selection strategies which relied instead on setting-specific sources of information (e.g., the student's level of knowledge and their speed of learning). By hypothesizing precisely how people's ostensibly suboptimal role-model-selection strategies may actually be potentially ecologically rational, our model adds to cultural evolutionary theory's understanding of its hypothesized on-average selective social learning.

To corroborate the hypothesis that humans achieved on-average selective social learning even for high-variance-payoff tasks, our work highlights the importance of identifying and investigating the relevant mechanisms of selective social learning, which would need to be robust in the face of high ecological fitness costs of overcommitting attention. One such mechanism, hypothesized by our model, is the potential dependence of task/role-model turnover strategies on setting-specific information, which can inform turnover even in the absence of retained environmental feedback. Another example of such a mechanism is the conformist or reputation-based nature of human role-model-selection strategies (Cavalli-Sforza & Feldman, 1981; Boyd & Richerson, 1985; Henrich, 2009). To illustrate, descriptive human role-model-selections rely at least partially on granting prestige status to role models based on popularity rather than on the relevant environmental feedback (Henrich & Gil-White, 2001).

These two mechanisms—reliance on setting-specific information and conformist role-model-selection strategies—are not competing explanations for on-average selective social learning in settings of high-variance environmental feedback. In fact, the latter mechanism may require the former, because in order for conformist role-model-selection strategies to facilitate selective social learning in the absence of environmental feedback, the prestige status granted to a popular role model may need to have had incorporated other helpful information at some point in the past. If this information could not feasibly have been environmental feedback, then it must have been setting-specific information in the complement of environmental feedback. Our theory proposes that the student's level of knowledge and their speed of learning can provide such setting-specific information to achieve an on-average selective strategy of task/role-model choice, even when retaining environmental feedback is unfeasible.

Regardless of whether our model is a good model of ancestral humans' learning environment, our test for verifying whether a learner is meaningfully incorporating their environmental observations into their decision-making—in the sense of classical Bayesian inference—may be general enough to have various potential applications. To illustrate, public-policy plans are often aimed at least partially at improving societal well-being. Arguably, the dominant paradigm with which this goal is approached is the assumption that each person's decisions (e.g., the price they are willing to pay or take for an item) reveal an informationally rational aggregate of their private observations relevant to their well-being (e.g., Harberger, 1971). Policymakers thus aim to economize on the cost of gathering copious, potentially idiosyncratic information by relying on each person's purported aggregate of their individual observations encapsulated by their decisions. The reliability of this information-gathering strategy is determined by whether each person is actually aggregating their observations in an informationally rational way.

However, as we have seen above, an extensive body of empirical evidence suggests that this assumption of informational rationality may not hold true when the relevant observations occur with high variance. Moreover, we have demonstrated the plausible ecological rationality of empirically robust cognitive biases by constructing an evolutionary model of social learning of task-specific knowledge, hypothesized by cultural evolutionary theory to be the primary mode of ancestral human learning. Our work thus contributes to raising the following research question: in which situations do public-policy plans aimed at improving societal well-being under the assumption of people's informational rationality actually succeed in doing so? It also begs a potentially important follow-up question: can public-policy plans be improved by replacing the assumption of informational rationality with the more empirically tenable assumption of ecological rationality? Domains of high-variance payoff data, such as gambling, may potentially be better served by the latter assumption over the former.

Another preliminary point is that informational rationality may not be an unattainable goal for human cognition. The decision-making of a person who is both trained in statistical methods and has the habit of applying this training to their own observations may be informationally rational. It may thus be fruitful not only to question the default assumption of people's informational rationality, but also to explore the potential upside of practical statistics training: such as the habit of keeping track of the mean past payoff data, as implied by our test for informational rationality. This statistical skill can be both a possible remedy to the potentially detrimental misassumption of informational rationality, and a facilitator of improved judgement and role-model selection at the individual level. One potential such benefit is dissuading people from socially learning the practice of repeated gambling on negative-expected-value lotteries.

4.2 Model limitations and directions for generalization

Our model is almost certainly an oversimplification of descriptive social task-learning, which in general involves extremely complex social dynamics. We non-exhaustively list several ways in which this is the case. We also note potential remedies, in the form of potential directions for generalization. Thereby generalizing our model may potentially enable it to better represent descriptive social task-learning and thereby better explain the relevant empirical data. We thus propose our model as a barebones representation of social, knowledge-based task learning, on which more sophisticated variants can potentially be built in the future (assuming, of course, that the thrust of the model's story is essentially correct).

First, our model's conclusion that the student retains no information from payoff data is oversimplified. Realistically, people can plausibly retain easy-to-remember aspects of their past payoff data, which may include the maximum and minimum payoff values observed so far. People may also temporarily retain a small number of recent payoff data, even when they fail to draw on more distant past data that a Bayesian-updating belief would incorporate. The realistic assumption that a small number of recent payoff observations may inform decision-making can account for additional empirically documented patterns in descriptive human learning, such as reinforcement learning (Nax & Perc, 2015).

Also, our model's assumption that knowledge affects decision-making through a unidimensional quantification—the level of knowledge b—is an oversimplification. There is no reason to believe that knowledge is unidimensional, an assumption we have used for the sole sake of tractably showing the evolutionary plausibility of recurrently non-monotonic confidence. In fact, given the sheer multifaceted nature of knowledge, we hypothesize that knowledge in general should affect decision-making through a more faithful, multidimensional quantification.

Moreover, our model's two-dimensional spectrum of task packages—assumed in our model to be comprised of a unidimensional knowledge-based difficulty level and a binary learning type—is an oversimplification. First, as we have noted above, knowledge is likely experienced as a multidimensional quantity, which makes it likely that a unidimensional knowledge-based classification of tasks is an oversimplification. Second, when a student attempts to learn from a role model, their method of learning would in general be placed somewhere on the spectrum between full imitation and full innovation. Third, our two-dimensional spectrum is unlikely to capture all the relevant variations in the task-learning process; idiosyncrasies of the task itself, of how the student learns, of how the teacher imparts (or ostensibly imparts) knowledge, and of the degree to which learning is student-directed as opposed to role-model-directed (for example, whether the student seeks out the role model for a task they already had in mind) may also influence the learning process. In particular, potentially consequential quantities like the speed of learning may vary with respect to characteristics of the task package that are not captured by this two-dimensional parametrization.

Finally, our model's assumption that task packages are drawn i.i.d. from a fixed probability distribution is an oversimplification. For one thing, the i.i.d. assumption on our model—added for the sake of tractability—ignores the likely correlations between different task packages due to similarities in either the teachers or the underlying tasks. For another, descriptive selection of tasks/role models is not well-modeled by an i.i.d. draw from a fixed distribution; it is better described as an intrinsically social process that involves dynamically occurring interactions between other students and other potential role models, such as via conformist role-model selection strategies (e.g., prestige status). Such a multi-agent interaction would need to be modeled by a complex game-theoretic model, rather than a comparatively tractable Bayesian decision-theoretic model (which can be solved by dynamic-programming-type methods under quite non-restrictive conditions). Regardless, only a model in the former formulation could veridically represent the relevant social dynamics, such as coordination and punishment.

4.3 Empirical tests

We sketch an empirical program to study descriptive human learning in the formulation of our theory. One of the primary goals of such a program would be the eventual corroboration or falsification of the theory itself. However, the program—by pursuing theoretical formulation—may also potentially yield other advances in the psychological sciences' understanding of descriptive human learning and decision-making, especially since the field is arguably held back by a shortage of theoretical formulation at the moment (Muthukrishna & Henrich, 2019).

First, we propose the empirical estimation of the true parameters of various social task-learning environments. Several parameters which we have proposed to be evolutionarily relevant include the proportion of attempted imitation that is successful, the proportion of unlearnable tasks among those that are learned by unsuccessful imitation (*de facto* innovation learning), the speed of each type of learning, ecological fitness costs of various action choices, and the situation-specific marginal payoff from a task. Empirical studies of how these parameters varied across both ancestral and contemporary human learning environments, as well as studies of whether they can predict the respective evolution of task-specific confidence and strategies of task/role-model turnover, would potentially contribute to a more robust and granular understanding of human cognition. Such studies would also allow us to test whether our model can veridically represent ancestral and contemporary human learning environments.

Estimates of such model parameters in ancestral environments would often be necessarily crude, given the general lack of archaeological and other relevant forms of evidence. As a start, one may feasibly expect ancestral humans who lived in areas where food is complicated to obtain (e.g., tundra)—when compared to those who lived in areas with easy food availability (e.g., rainforests)—to either have a generally lower-valued payoff function, a task difficulty distribution biased towards higher difficulty values, or a greater probability of unlearnable tasks. Empirical studies can then test whether these hypothesized parameter differences in the ancestral environments affect strategies of task-payoff estimation and task/role-model turnover in the ways predicted by our model, such as Proposition 1's prediction that task-specific confidence (conditional on learning not yet having completed) decreases in the proportion of unlearnable tasks. Such efforts, however, may be inevitably limited, due to the multitude and granular variation of the model parameters, the difficulty of measuring many of them for ancestral environments, and the uncertainty in whether ecologically rational social task-learning strategies were selected via genetic evolution.

More immediately promising would be applying such efforts to investigating the social task-learning of evolutionarily relevant foragers whose lifestyles are hypothesized to be faithful continuations of their ancestors', such as the Hadza people (Marlowe, 2010; Lew-Levy et al., 2021). Such efforts will not be confounded by our current uncertainty in whether the adaptive mechanism by which ecologically rational social task-learning strategies were selected was genetic evolution or contemporary learning. We propose empirical studies of the social task-learning of such peoples as a potentially fruitful first step in testing whether our model (or a sufficient generalization) is a good model of descriptive human learning. If the answer to this question is affirmative, empirical researchers can proceed to study learning environments with granular variations in model parameters, genetic-evolutionary background, and cultural-evolutionary background. Doing so may further corroborate or potentially falsify our model, determine the role of genetic evolution and contemporary learning in the selection of its ecological rational strategies, and investigate the scientific consequences of any such findings.

For instance, suppose that our model is a good model of descriptive human learning, and that the adaptive mechanism by which its ecologically rational strategies were selected was at least partially genetic evolution. Then, our model may provide a way in which otherwise mysterious aspects of ancestral human learning environments can be studied indirectly: via empirical studies (of task-specific confidence and task/role-model turnover) investigating people living today. Specifically, empirical data of these psychological aspects—which are comparatively easy to obtain—can narrow down the feasible region of model parametrizations that can evolutionarily explain the data of such studies. This would then potentially inform us of characteristics of the respective ancestral human learning environments that would otherwise be difficult to discern. On the other hand, suppose that the adaptive mechanism by which the model's ecologically rational strategies were selected was at least partially contemporary cultural learning. Then, our model may similarly enable certain aspects of a cultural group's social task-learning environment to yield consequences about certain aspects of their decision-making, and vice versa. Such a bridge between different objects of study can increase the number of ways we can study each, and thereby contribute to a more comprehensive literature on human cognition.

It is evident that in all lines of inquiry described above, empirical data from contemporary people's learning (including, but not limited to social task-specific learning) could be crucial. Such data can be obtained from lab studies and field studies of the relevant psychological aspects. A prediction of our theory is that these psychological aspects may be evolutionarily affected by independent variables that are specific to social, knowledge-based learning and not to individual learning: even when in ostensibly unambiguous settings of individual learning with costless environmental feedback. Therefore, it may be potentially beneficial for empirical studies of these psychological aspects—even in domains of individual learning—to keep track of potentially social-learning-specific independent variables like the level of knowledge, the speed of learning, and whether the method of learning is imitation or innovation.

Another prediction is that two psychological aspects in particular—task-specific confidence and task/role-model turnover—are evolutionarily related. We thus propose that they should be studied concurrently. In particular, empirical studies should look for our theory's hypothesized, potentially discernable piecewise cutoff point (a "phase transition") in the student's task-specific confidence, which should exist and coincide with the identification of the learning type. They should then investigate precisely when this cutoff point—as well as task/role-model turnover—occurs, which should vary with respect to whether the student's learning method is authentic imitation or *de facto* innovation in ways that are elucidated by our model.

Lab studies would do well to incorporate the excellent experimental design of Sanchez and Dunning (2018, 2020), which is effective at studying task-specific confidence over the course of learning a high-variance-outcome task over repeated attempts. To arrive at the setting of our model, the Sanchez–Dunning experimental design could be modified to represent an unambiguous setting of task-specific learning via attempted imitation. Ideally, this modified design would achieve a dichotomy between successful imitation and *de facto* innovation (e.g., by having some role models teach via the Socratic method, and other role models provide actually helpful knowledge: but not to the point of trivializing task learning), include unlearnable tasks (e.g., by having payoffs of unlearnable tasks occur with full randomness that cannot ever be predicted), grant the option of drawing a new task and/or role model, and—just as in the original experiment—offer an incentive-compatible reward. Such an experimental design could then essentially be a parametrization of our learning environment, albeit an artificial one and not an ancestral one. These artificial model parameters, the genetic and cultural-evolutionary background of the experimental subjects, and other potentially relevant treatment effects (independent variables) can then be varied across studies to test the quantitative predictions of our theory regarding task-specific confidence and task/role-model turnover.

Also, on top of such an artificial model parametrization, empirical researchers could add other hypothesized cultural-evolutionary-theoretic mechanisms that would endow its learning environment with an unambiguously social context. Key examples of such mechanisms include a nontrivial amount of choice in the selection of new tasks and/or role models, the ability to observe the number of other students that have chosen each task/role model, and the ability to exchange information with other students and role models. The inclusion and veridical representation of such mechanisms could be key to investigating cultural-evolutionary-theoretic dynamics that are not fully captured in a decision-theoretic setting such as that of our model.

In addition, empirical researchers could pursue field research of social task-specific learning, especially pertaining to tasks with high-variance payoffs. In contrast to the lab research proposed above, field research would allow for a more veridical representation of social task-specific learning, at the potential expense of experimental controls and granular variation of the independent variables. Doing such field studies in a manner that comprehensively measures all data relevant to our model would be undoubtedly challenging, given that it may need to keep track of every student and role model's interactions, respective levels of experience, respective speeds of learning, respective payoff data, and—if technologically feasible—informative measurements of knowledge. Even if all such data were collected, there may additionally need to be some degree of nontrivial inference from the data to discern certain model parameters: for example, which packages of tasks and role models were learned via successful imitation rather than *de facto* innovation. Future advances towards improving and widening the collected data in such field studies would potentially help on these fronts.

In both field studies and lab studies investigating descriptive social learning of high-variance-payoff tasks, empirical researchers would do well to take into account the sheer diversity in potential subjects' psychological profiles and treatment effects, which should ideally be recorded as comprehensively as possible in order to keep track of all potential independent variables (Yarkoni, 2020). In fact, consider the following two hypotheses. First, subjects who are most likely to be studied by lab research—individuals of Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies—are in important ways psychological outliers relative to the rest of the human population (Henrich et al., 2010). Second, much of the genus *homo's* two-million-year existence was spent in the non-WEIRD lifestyle of mobile foragers (Townsend, 2018). A consequence is that a comprehensive understanding of descriptive human learning may require studying the social task-learning of mobile foragers whose lifestyles are faithful continuations of their ancestors': and studying that of non-WEIRD peoples in general. To their credit, field studies are already doing so extensively (e.g., Kline et al., 2013; Lew-Levy et al., 2017, 2021; Lew-Levy & Boyette, 2018; Salali et al., 2019, 2016; Schniter et al., 2015). It may potentially be fruitful to have more of the relevant lab studies, such as the Sanchez–Dunning experimental design (2018, 2020), to also be targeted at individuals of non-WEIRD societies.

Empirical tests of our model's assumptions themselves would be potentially valuable for the purpose of assessing whether it is a good model of ancestral learning environments. The program to investigate whether social task-learning comprised the primary selection pressure of ancestral human learning is not new. It is a vibrant line of inquiry that constitutes the center of the debate between cultural evolutionary theory and its competing hypotheses (Baimel et al., 2021), whose resolution has potential implications for other debates: like that regarding the hypothesized evolution of moral, norm-based preferences (Capraro & Perc, 2021). Our model contributes new insights that can add to this program. Most notably, it demonstrates that cultural evolutionary theory can explain otherwise puzzling cognitive biases like recurrently non-monotonic confidence. The fact that descriptive human learning is thereby cognitively biased—even in unambiguous settings of individual learning with costless environmental feedback—grants plausibility to cultural evolutionary theory's hypothesis that the primary selection pressure on ancestral human cognition was social, knowledge-based task-learning.

Also, our model identifies several potentially relevant mechanisms in a hypothesized learning environment of social, knowledge-based task-learning: for example, the classification of attempted imitation learning into successful imitation and *de facto* imitation learning, as well as the risk of an unlearnable task in the case of the latter. In particular, it explicitly posits the predictive importance of ecological fitness costs of overcommitting attention, which determine whether the evolutionarily optimal strategy of selective role-model selection meaningfully learns from the relevant payoff data. Our model's formalization of these parameters can augment empirical assessments of cultural evolutionary theory by informing a potentially fruitful avenue of research: specifically, the estimation of these parameters for various, potentially ancestral learning environments; combined with an investigation of cultural evolutionary theory's relevant predictions and of the degree to which these predictions hold. One such prediction from our model (and suitable generalizations of it) would be that when the ecological fitness cost of retaining payoff data is sufficiently high, the optimal strategy of task/role-model turnover would not retain it, and instead rely on other sources of information that are specific to the hypothesized setting of social, knowledge-based task-learning.

A stronger claim of our theory is that the costly cognitive mechanism by which ancestral humans distinguished successful imitation from *de facto* innovation was a mental time-measurement experiment, to distinguish their respective learning speeds. Our hypothesized existence of such mental time-measurement experiments is a special case of the generally theorized mental evidence-sampling process preceding a decision (e.g., Pleskac & Busemeyer, 2010). Empirical tests of our assumption that the speed of imitation is faster than that of innovation, as well as of our assumption that human learners can and do differentiate between the two speeds via mental time-measurement, could help probe the plausibility of our theory.

Other plausible hypotheses for the cognitive mechanism by which the student differentiates between imitation and innovation include a costly-to-observe signal effused by the teacher, or one effused by the accumulated task-specific knowledge at any given point of time. Our model can be suitably modified to use such an alternative hypothesis for this cognitive mechanism. In fact, incorporating such an alternative hypothesis would make the model considerably simpler, since it would not need to consider variation in learning speeds. However, a disadvantage of such an alternative hypothesis is that empirically testing it may be less straightforward, at least without relying on neuroscientific methods. We have thus not pursued these alternatively hypothesized mechanisms in the present paper, although we do not rule their veracity out and hope that they may be feasibly testable in the future.

More generally, it may be plausible that future developments in our neuroscientific knowledge will enable a detailed mechanistic understanding of descriptive human learning. While remarkable empirical advances have been made on this front, our current level of neuroscientific understanding has a long way to go, given the extreme complexity of human cognition and the relative adolescence of the field of neuroscience. However, our sketches of potential empirical studies demonstrate that even at our currently limited level of understanding of descriptive human learning, substantive progress—towards testing our theory and in general—may be plausible. Moreover, evolutionary-theoretic hypotheses like those of our model can inform the design, data collection, and analyses of such empirical studies, and thereby partially compensate for the preliminary nature of the current neuroscientific literature. Given the immediate and far-reaching upside of a comprehensive understanding of descriptive human decision-making, we propose that the eventual benefits of a cumulative program of research working towards this goal (even prior to a full neuroscientific understanding) may outweigh the costs.

Acknowledgements

The author thanks Eric Maskin for suggesting the problem and for giving invaluable guidance throughout the course of this work. The author also thanks Joe Henrich for sharing his immense expertise, invaluable advice on the manuscript, and constant encouragement, all of which have been indispensable for this work. Moreover, the author thanks Chris Bakerlee, Rahul Bhui, Matt Cashman, Charles Efferson, Holly Elmore, Ben Golub, Mark Kisin, Philipp Schoenegger, Suzanne Smith, Alex Stewart, Tomasz Strzalecki, Suren Tavakalov, Wesley Wildman, and Alice Zhu for providing many helpful comments on this paper. Additionally, the author thanks members of the Henrich lab for a substantially helpful feedback session. In particular, he is grateful to Matt Cashman for introducing him to the work of Gerd Gigerenzer, Cammie Curtin for introducing him to the work of Charles Efferson, Patricia Greenfield for providing practical advice on the revision process, Ivan Kroupin for suggesting an emphasis on social learning, Graham Noblit for providing detailed feedback on the manuscript, Rachna Reddy for suggesting an emphasis on concrete ecological fitness costs, and TC Zeng for providing detailed feedback on the abstract and on the topic of selective social learning. Finally, the author thanks Joe Henrich for introducing him to the references Henrich, 2009; Muthukrishna et al., 2018, and Singh, 2018; Barbara Finlay for introducing him to Gladwell, 2019 and Yarkoni, 2020; and Will MacAskill for introducing him to Henrich, 2015.

Funding

This work was supported by the National Science Foundation through the Graduate Research Fellowship Program (grant number DGE1745303), the Centre for Effective Altruism through the Global Priorities Fellowship, and Harvard University through graduate student fellowships.

References

Agarwal, S., Driscoll, J. C., Gabaix, X., & Laibson, D. (2008). *Learning in the credit card market* [Harvard University Working Paper]. doi: 10.3386/w13822

Akerlof, G. A., & Shiller, R. J. (2009). *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism.* Princeton, NJ: Princeton University Press.

Augenblick, N., & Rabin, M. (2021). Belief movement, uncertainty reduction, and rational updating. *Quarterly Journal of Economics*, *136*(2), 933–985. doi: 10.1093/qje/qjaa043

Baimel, A., Juda, M., Birch, S., & Henrich, J. (2021). Machiavellian strategist or cultural learner? Mentalizing and learning over development in a resource-sharing game. *Evolutionary Human Sciences*, *3*. doi: 10.1017/ehs.2021.11

Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics*, *116*(1), 261–292. doi: 10.1162/003355301556400

Becker, G., Degroot, M., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, *9*(3), 226–232. doi: 10.1002/bs.3830090304

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics—Foundations and Applications 2* (pp. 69–186). North-Holland, Amsterdam. doi: 10.1016/bs.hesbe.2018.11.002

Bird-David, N. (1992). Beyond 'The hunting and gathering mode of subsistence': Culture-sensitive observations on the Nayaka and other modern hunter-gatherers. *Man*, *27*(1), 19–44. doi: 10.2307/2803593

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, R., & Richerson, P. J. (1988). An evolutionary model of social learning: The effects of spatial and temporal variation. In T. Zentall & B. Galef (Eds.), *Social Learning: Psychological and Biological Perspectives* (pp. 29–48). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. doi: 10.4324/9781315801889

Boyd, R., & Richerson, P. J. (1995). Why does culture increase human adaptability? *Ethology and Sociobiology*, *16*(2), 125–143. doi: 10.1016/0162-3095(94)00073-G

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77. doi: 10.1037/0022-3514.90.1.60

Capraro, V., & Perc, M. (2021). Mathematical foundations of moral preferences. *Journal of the Royal Society interface*, *18*(175), 20200880–20200880. doi: 10.1098/rsif.2020.0880

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, N.J.: Princeton University Press.

Corner, A., & Hahn, U. (2012). Normative theories of argumentation: Are some norms better than others? *Synthese*, *190*(16), 3579–3610. doi: 10.1007/s11229-012-0211-y

de Francesco, G. (1939). *The Power of the Charlatan*. New Haven, CT: Yale University Press.

DellaVigna, S., & Malmendier, U. (2006). Paying not to go to the gym. *American Economic Review*, *96*(3), 694–719. doi: 10.1257/aer.96.3.694

Dixon, N. F. (1976). *On the Psychology of Military Incompetence*. London: Cape.

Doob, J. L. (1949). Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin – 3 Juillet, 1948),* (pp. 23–27). Paris: Paris CNRS.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. doi: 10.1016/j.obhdp.2007.05.002

Freedman, D. A. (1963). On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Annals of Mathematical Statistics*, *34*(4), 1386–1403.

Galanter, M. (1989). *Cults: Faith, Healing, and Coercion*. New York: Oxford University Press.

Gigerenzer, G. (2000). *Adaptive Thinking: Rationality in the Real World*. Oxford; New York: Oxford University Press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple Heuristics that Make Us Smart*. New York: Oxford University Press.

Gladwell, M. (2019). *Talking to Strangers: What We Should Know about the People We Don't Know*. London: Allen Lane.

Harberger, A. C. (1971). Three basic postulates for applied welfare economics: An interpretive essay. *Journal of Economic Literature*, *9*(3), 785–797.

Haselhuhn, M. P., Pope, D. G., Schweitzer, M. E., & Fishman, P. (2012). The impact of personal experience on behavior: Evidence from video-rental fines. *Management Science*, *58*(1), 52–61. doi: 10.1287/mnsc.1110.1367

Haun, D. E., Zeringue, A., Leach, A., & Foley, A. (2000). Assessing the competence of specimen-processing personnel. *Laboratory Medicine*, *31*(11), 633–637. doi: 10.1309/8Y66-NCN2-J8NH-U66R

Henrich, J. (2009). The evolution of costly displays, cooperation and religion: credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, *30*(4), 244–260. doi: 10.1016/j.evolhumbehav.2009.03.005

Henrich, J. (2015). *The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter.* Princeton: Princeton University Press.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196. doi: 10.1016/S1090-5138(00)00071-4

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Hoffman, M., & Burks, S. V. (2020). Worker overconfidence: Field evidence and implications for employee turnover and firm profits. *Quantitative Economics*, *11*(1), 315–348. doi: 10.3982/QE834

Hooper, P. L., Demps, K., Gurven, M., Gerkey, D., & Kaplan, H. S. (2015). Skills, division of labour and economies of scale among Amazonian hunters and South Indian honey collectors. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *370*(1683), 20150008. doi: 10.1098/rstb.2015.0008

Humphrey, N. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing Points in Ethology.* Cambridge, UK: Cambridge University Press.

Insler, M., McQuoid, A. F., Rahman, A., & Smith, K. A. (2021). Fear and loathing in the classroom: Why does teacher quality matter? *IZA Discussion Paper No. 14036*. Retrieved from `https://ssrn.com/abstract=3767273`

Jansen, R. A., Rafferty, A. N., & Griffiths, T. L. (2021). A rational model of the Dunning–Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*. doi: 10.1038/s41562-021-01057-0

Johnson, D. D. P. (2004). *Overconfidence and War: The Havoc and Glory of Positive Illusions*. Cambridge, MA: Harvard University Press.

Jones, H., Diop, N., Askew, I., & Kabore, I. (1999). Female genital cutting practices in Burkina

Faso and Mali and their negative health outcomes. *Studies in Family Planning*, *30*(3), 219–230. doi: 10.1111/j.1728-4465.1999.00219.x

Kline, M. A., Boyd, R., & Henrich, J. (2013). Teaching and the life history of cultural transmission in Fijian villages. *Human Nature*, *24*(4), 351–374. doi: 10.1007/s12110-013-9180-1

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of It: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Laland, K. N. (2017). *Darwin's Unfinished Symphony: How Culture Made the Human Mind*. Princeton: Princeton University Press.

Lew-Levy, S., & Boyette, A. H. (2018). Evidence for the adaptive learning function of work and work-themed play among Aka forager and Ngandu farmer children from the Congo Basin. *Human Nature*, *29*(2), 157–185. doi: 10.1007/s12110-018-9314-6

Lew-Levy, S., Reckin, R., Lavi, N., Cristóbal-Azkarate, J., & Ellis-Davies, K. (2017). How do hunter-gatherer children learn subsistence skills? A meta-ethnographic review. *Human Nature*, *28*, 367–394. doi: 10.1007/s12110-017-9302-2

Lew-Levy, S., Ringen, E. J., Crittenden, A. N., Mabulla, I. A., Broesch, T., & Kline, M. A. (2021). The life history of learning subsistence skills among Hadza and BaYaka foragers from Tanzania and the Republic of Congo. *Human Nature*, *32*(1), 16–47. doi: 10.1007/s12110-021-09386-9

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159 – 183. doi: 10.1016/0030-5073(77)90001-0

Lindenbaum, S. (2001). Kuru, prions, and human affairs: Thinking about epidemics. *Annual Review of Anthropology*, *30*(1), 363–385. doi: 10.1146/annurev.anthro.30.1.363

Lisi, M., Mongillo, G., Milne, G., Dekker, T., & Gorea, A. (2021). Discrete confidence levels revealed by sequential decisions. *Nature Human Behaviour*, *5*(2), 273–280. doi: 10.1038/s41562-020-00953-1

MacKenzie, D. (1994). Computer-related accidental death: an empirical exploration. *Science and Public Policy*, *21*(4), 233–248. doi: 10.1093/spp/21.4.233

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. *Journal of Finance*, *60*(6), 2661–2700. doi: 10.1111/j.1540-6261.2005.00813.x

Marlowe, F. (2010). *The Hadza: Hunter-gatherers of Tanzania*. Berkeley: University of

California Press.

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, *31*(5), 309–319. doi: 10.1016/j.evolhumbehav.2010.04.005

McNamara, J., & Houston, A. (1980). The application of statistical decision theory to animal behaviour. *Journal of Theoretical Biology*, *85*(4), 673–690. doi: 10.1016/0022-5193(80) 90265-9

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. doi: 10.1037/0033-295X.115.2.502

Muthukrishna, M., Doebeli, M., Chudek, M., & Henrich, J. (2018). The Cultural Brain Hypothesis: How culture drives brain expansion, sociality, and life history. *PLoS Computational Biology*, *14*(11), e1006504–e1006504. doi: 10.1371/journal.pcbi.1006504

Muthukrishna, M., & Henrich, J. (2016). Innovation in the collective brain. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *371*(1690), 20150192. doi: 10.1098/rstb.2015.0192

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. doi: 10.1038/s41562-018-0522-1

Nax, H. H., & Perc, M. (2015). Directional learning and the provisioning of public goods. *Scientific Reports*, *5*(1), 8010–8010. doi: 10.1038/srep08010

Park, Y. J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. *Theory and Decision*, *69*(1), 143–166. doi: 10.1007/s11238-010-9200-0

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, *107*(Supplement 2), 8993–8999. doi: 10.1073/pnas.0914630107

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. doi: 10.1037/a0019737

Reader, S. M., Hager, Y., & Laland, K. N. (2011). The evolution of primate general and cultural intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1017–1027. doi: 10.1098/rstb.2010.0342

Salali, G. D., Chaudhary, N., Bouer, J., Thompson, J., Vinicius, L., & Migliano, A. B. (2019). Development of social learning and play in BaYaka hunter-gatherers of Congo. *Scientific Reports*, *9*(1), 11080–10. doi: 10.1038/s41598-019-47515-8

Salali, G. D., Chaudhary, N., Thompson, J., Grace, O. M., van der Burgt, X. M., Dyble, M.,

… Migliano, A. B. (2016). Knowledge-sharing networks in hunter-gatherers and the evolution of cumulative culture. *Current Biology*, *26*(18), 2516–2521. doi: 10.1016/j.cub.2016.07.015

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. doi: 10.1037/pspa0000102

Sanchez, C., & Dunning, D. (2020). Decision fluency and overconfidence among beginners. *Decision*, *7*(3), 225–237. doi: 10.1037/dec0000122

Savage, L. J. (1972). *The Foundations of Statistics* (2d rev. ed. ed.). New York: Dover Publications.

Scheinkman, J. A., & Xiong, W. (2003). Overconfidence and speculative bubbles. *Journal of Political Economy*, *111*(6), 1183–1220. doi: 10.1086/378531

Scheirer, W. (2020). A pandemic of bad science. *Bulletin of the Atomic Scientists*, *76*(4), 175–184. doi: 10.1080/00963402.2020.1778361

Schlosser, E. (2013). *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*. New York: The Penguin Press.

Schniter, E., Gurven, M., Kaplan, H. S., Wilcox, N. T., & Hooper, P. L. (2015). Skill ontogeny among Tsimane forager-horticulturalists. *American Journal of Physical Anthropology*, *158*(1), 3–18. doi: 10.1002/ajpa.22757

Singh, M. (2018). The cultural evolution of shamanism. *Behavioral and Brain Sciences*, *41*, 1–83. doi: 10.1017/S0140525X17001893

Street, S. E., Navarrete, A. F., Reader, S. M., & Laland, K. N. (2017). Coevolution of cultural intelligence, extended life history, sociality, and brain size in primates. *Proceedings of the National Academy of Sciences*, *114*(30), 7908–7914. doi: 10.1073/pnas.1620734114

Townsend, C. (2018). *Egalitarianism, evolution of.* Oxford: John Wiley and Sons.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. doi: 10.1126/science.185.4157.1124

Uscinski, J. E., Klofstad, C., & Atkinson, M. D. (2016). What drives conspiratorial beliefs? The role of informational cues and predispositions. *Political Research Quarterly*, *69*(1), 57–71. doi: 10.1177/1065912915621621

Valone, T. J. (2006). Are animals capable of Bayesian updating? An empirical review. *Oikos*, *112*(2), 252–259. doi: 10.1111/j.0030-1299.2006.13465.x

van Schaik, C. P., & Burkart, J. M. (2011). Social learning and evolution: the cultural intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1008–1016. doi: 10.1098/rstb.2010.0304

Wagner, N. (2015). Female genital cutting and long-term health consequences – Nationally representative estimates across 13 countries. *The Journal of Development Studies*, *51*(3), 226–246. doi: 10.1080/00220388.2014.976620

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *Journal of Economic Education*, *40*(3), 227–261. doi: 10.3200/JECE.40.3.227-261

Whiten, A., & van Schaik, C. P. (2007). The evolution of animal 'cultures' and social intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 603–620. doi: 10.1098/rstb.2006.1998

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. doi: 10.1017/S0140525X20001685

Model parameters of the modified Bayesian model

1. the marginal payoff distribution $\varphi(a,b) \in \mathcal{P}(S)$ and its expected value $f(a,b)$, for every $a \in (0,\infty) \cup \{\infty\}$ and finite $0 \leq b \leq a$,
2. the discrete knowledge jumps $\{B(i) : i \in \mathbb{N}, i > 0\}$,
3. the learning period lengths $\{\Delta_j(i) : i \in \mathbb{N}, i > 0\}$ for $j \in \{im, in\}$,
4. the exponential discount factor δ of time,
5. the proportion p of infinite-difficulty tasks among all innovation-learning tasks,
6. the proportion q of imitation-learning tasks among all tasks,
7. the exponential discount factor η of the distribution of task difficulty values,
8. the fraction of time r of task attempts that can be devoted to alternative foraging opportunities,
9. the distribution $\psi \in \mathcal{P}(S)$ of the marginal payoffs of alternative foraging opportunities,
10. the expected cost $-C_{\text{retain}}$ of retaining a payoff observation, and
11. the expected cost $-C_{\text{identify}}$ of a mental time-measurement experiment to identify the learning type j.

Table 1. List of model parameters of the Bayesian model modified to represent ancestral human learning, presented in Subsection 2.2. These model parameters are required to satisfy the conditions discussed earlier in this subsection.

Algorithmic description of the modified Bayesian model

1. The student draws from the distribution μ the task (j, a), the value of which is unknown to them. The attempt number specific to the task, i, is set to zero, and their level of knowledge b is set to zero. The time value T is set to zero.
2. The student carries out the ith attempt of the current task, which constitutes the following.
   - First, the student draws from the distribution ψ a random marginal payoff $s \in S$ whose value is known to them, and decide whether to forgo a fraction r of the task attempt for this alternative marginal payoff.
   - Second, the student decide whether to pay an expected cost $-C_{\text{identify}}$ for a time-measurement experiment to identify $\Delta_j(i)$, which is only possible if $\Delta_{im}(i) < \Delta_{in}(i)$ rather than $\Delta_{im}(i) = \Delta_{in}(i)$.
   - Third, they spend the time $\Delta_j(i)$ on the task attempt (T is incremented by this amount), at the end of which they receive a payoff of

     $$\begin{cases} \delta^T \left(rs + (1-r)\bar{s}\right) \int_0^{\Delta_j(i)} \delta^t dt & \text{if the student had decided to forgo,} \\ \delta^T \bar{s} \int_0^{\Delta_j(i)} \delta^t dt & \text{otherwise,} \end{cases} \tag{99}$$

     where $\bar{s} \in S$ is drawn from the distribution $\varphi(a, b)$. The student chooses whether to retain the observation $\bar{s}$ of the payoff value.
   - Fourth, if the student had performed a time-measurement experiment during this learning attempt, then they learn the value $\Delta_j(i)$ and thereby, the learning type j.
   - Fifth, b discretely jumps to the next level—$B(i+1)$ or a, whichever is smaller—and the index i is incremented by one.
   - Finally, the student chooses whether to quit the current task. If so, they draw a new task (j, a) from μ (independently with respect to the previously drawn tasks), b is set to zero, and i is set to zero. Otherwise, they continue to learn the task attempt at the new level of experience $i + 1$.
3. Step 2 is infinitely repeated.

Table 2. An algorithmic description of the Bayesian model modified to represent ancestral human learning, presented in Subsection 2.2.

Model parameters of the continuous learning model
1. the marginal payoff function $f(a, b)$,
2. the imitation-learning knowledge function $L_{im,\infty}(t)$,
3. the innovation-learning knowledge function $L_{in,\infty}(t)$,
4. the exponential discount factor δ of time,
5. the proportion p of infinite-difficulty tasks among all innovation-learning tasks,
6. the proportion q of imitation-learning tasks among all tasks,
7. the exponential discount factor η of the distribution of task difficulty values,
8. the constant β constraining the student's quitting.

Table 3. List of model parameters of the continuous learning model, which approximates our modified Bayesian model of ancestral human learning. The continuous learning model is presented in Subsection 3.2.

Algorithmic description of the continuous learning model

1. Time is set to $T = 0$.
2. The student draws from the distribution μ the task (j, a), the value of which is unknown to them.
3. If the student's quitting strategy is $\boldsymbol{b} = b$, then they receive a payoff of

$$\int_{T}^{T+L_{j,\infty}^{-1}(b)} \delta^t f(a, L_{j,\infty}(t))dt, \tag{100}$$

   and T is incremented by $L_{j,\infty}^{-1}(b)$. If the student's quitting strategy is $\boldsymbol{b} = (b_{im}, b_{in})$, then they receive a payoff of

$$\int_{T}^{T+L_{j,\infty}^{-1}(b_j)} \delta^t f(a, L_{j,\infty}(t))dt. \tag{101}$$

   and time is incremented by $L_{j,\infty}^{-1}(b_j)$.
4. If $T = \infty$, the algorithm is complete. If T is finite, return to Step 2 and repeat it along with the following steps.

Table 4. An algorithmic description of the continuous learning model, which approximates our modified Bayesian model of ancestral human learning. The continuous learning model is presented in Subsection 3.2.

Example family of model parametrizations $\boldsymbol{M}(n)$ of the modified Bayesian model

1. The time-discount factor is $\delta = 0.9$.
2. The marginal payoff function is $f(a, b) = b/a$.
3. The proportion of unlearnable tasks among those learned by innovation is $p = 0.01$.
4. The proportion of tasks that are learned by imitation is $q = 0.01$.
5. The decay factor of task difficulty values is $\eta = 0.5$.
6. The learning period lengths are given by

$$\Delta_{im}(i, n) = \begin{cases} \frac{2}{n+1} & \text{if } i < n+1, \\ \frac{1}{n+1} & \text{if } i \geq n+1, \end{cases} \tag{102}$$

and

$$\Delta_{in}(i, n) = \frac{2}{n+1}. \tag{103}$$

7. The knowledge jump values are given by $B(i, n) = \frac{2i}{n+1}$.
8. The expected cost of a time-measurement experiment to identify the learning type is $-C_{\text{identify}}$ for $C_{\text{identify}} = \frac{1}{n+1}$.
9. The fraction of time of task attempts that can be devoted to alternative foraging opportunities is given by $r = e^{-(n+1)}$.
10. The distribution ψ of the marginal payoffs of alternative foraging opportunities is arbitrary.
11. The distributions $\varphi(a, b)$ can be arbitrarily chosen, as long as we have $\mathbb{E}[\varphi(a, b)] = f(a, b)$.
12. As we have assumed throughout the paper, the expected cost of retaining a payoff observation, $-C_{\text{retain}}$, has sufficiently high magnitude C_{retain} so that payoff data are never retained: e.g., large enough so that the inequality (35) holds.

Table 5. An example family of parametrizations $\boldsymbol{M}(n)$ of our modified Bayesian model of ancestral human learning. The continuous learning model approximating this family, $\boldsymbol{M}(\infty)$, is characterized by a non-monotonic confidence function (see Figure 2). It follows that for sufficiently large n, the evolutionarily optimal confidence function of the model parametrization $\boldsymbol{M}(n)$ is also non-monotonic.

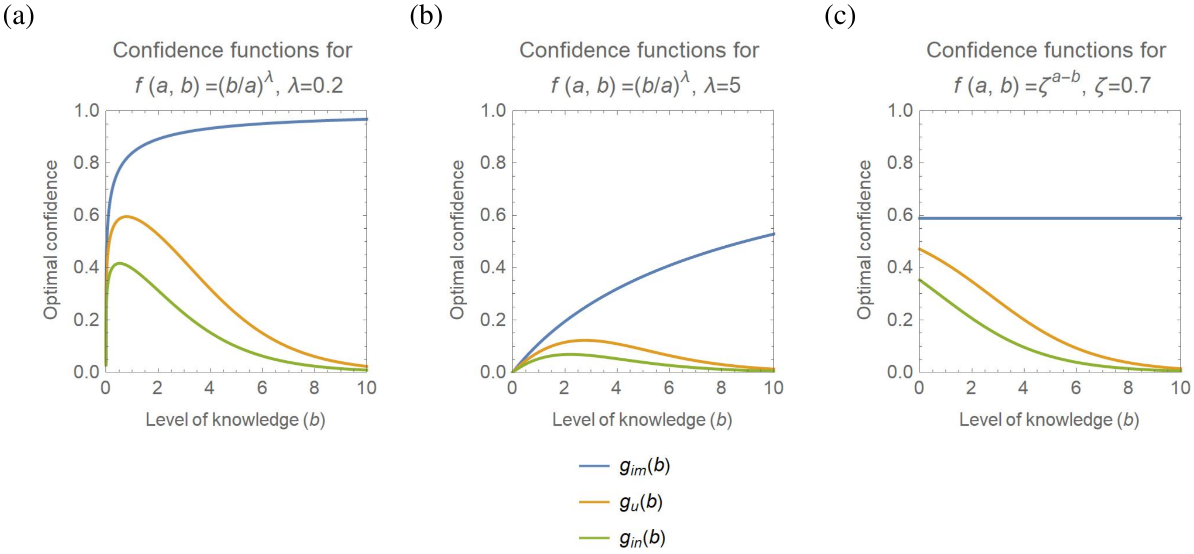

Figure 1. The imitation-learning confidence function $g_{im}(b)$, the innovation-learning confidence function $g_{in}(b)$, and the unconditional confidence function $g_u(b)$ for model parameter choices $p = 0.4, q = 0.5, \eta = 0.6$, and varying payoff function $f(a, b)$; note that the other model parameters do not affect these confidence functions. Consistent with Proposition 1, we have the inequalities $g_{in}(b) < g_u(b) < g_{im}(b)$. Also, consistent with Proposition 2(a), when the payoff function $f(a, b)$ satisfies Assumption 1—panels (a) and (b)—the imitation-learning confidence function $g_{im}(b)$ is strictly increasing. The payoff function of panel (c) does not satisfy Assumption 1. As a result, the corresponding imitation-learning confidence function $g_{im}(b)$ is not necessarily strictly increasing (in fact, it is constant). Finally, consistent with Proposition 2(b), the confidence functions $g_{in}(b)$ and $g_u(b)$ are eventually decaying to zero.

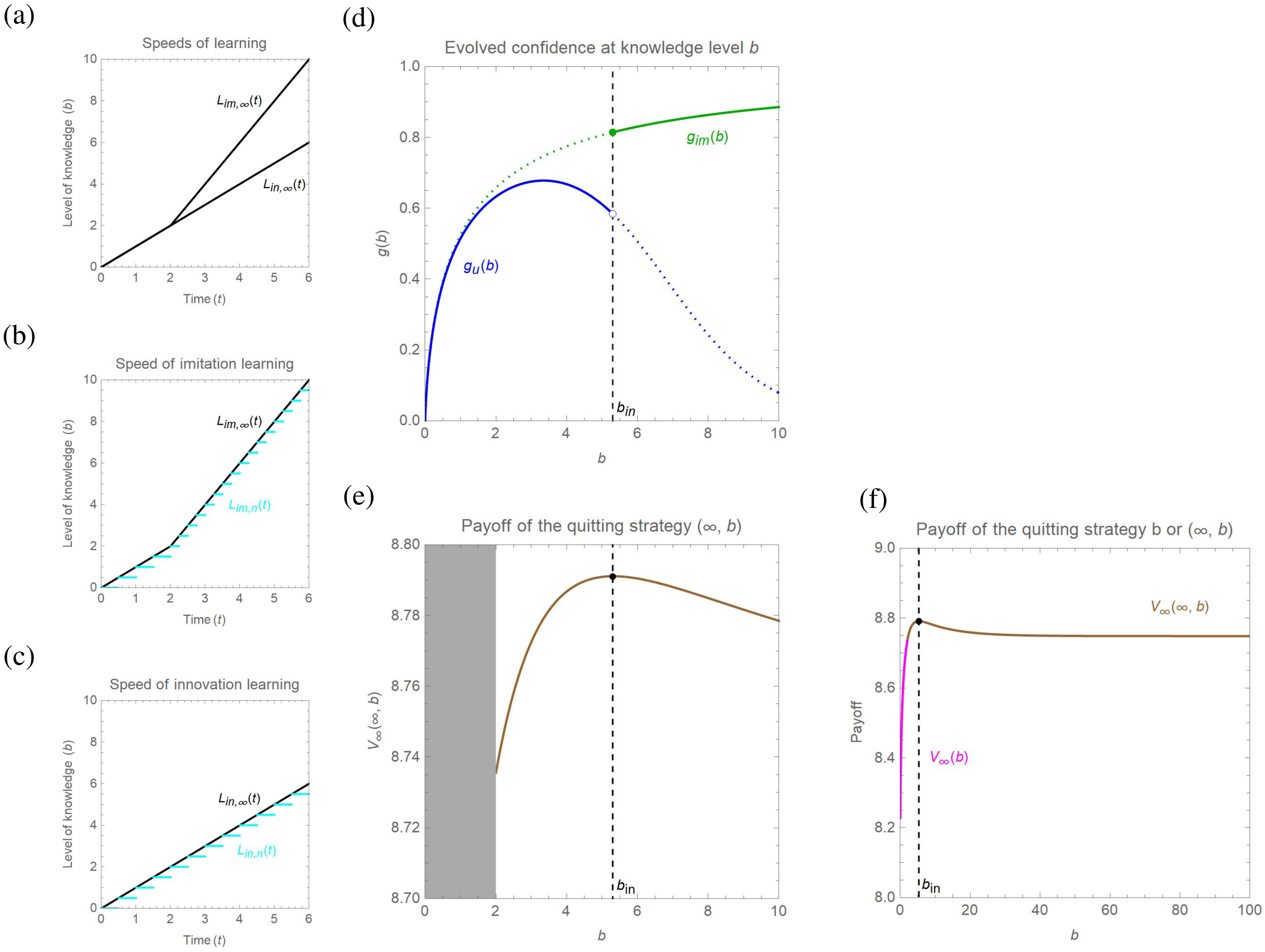

Figure 2. Plots of quantities relevant to the family of model parametrizations $\{\boldsymbol{M}(n)\}_{n\in\mathbb{N}}$ and the approximating continuous learning model $\boldsymbol{M}(\infty)$, presented in Table 5. Panel (a) plots the knowledge functions $L_{im,\infty}(t)$ and $L_{in,\infty}(t)$ of $\boldsymbol{M}(\infty)$, panel (b) shows how $L_{im,\infty}(t)$ approximates the imitation-learning knowledge functions $L_{im,n}(t)$ of $\boldsymbol{M}(n)$ ($n = 3$ is pictured), panel (c) shows how $L_{in,\infty}(t)$ approximates the imitation-learning knowledge functions $L_{in,n}(t)$ of $\boldsymbol{M}(n)$ ($n = 3$ is pictured), panel (d) plots the evolutionarily optimal estimate of confidence $g(b)$ (conditional on learning not yet having completed) in $\boldsymbol{M}(\infty)$, panel (e) plots the payoff $V_\infty(\boldsymbol{b})$ of the quitting strategy $\boldsymbol{b} = (\infty, b)$ for $b \geq \beta$, and panel (f) increases the domain and additionally plots the payoff $V_\infty(\boldsymbol{b})$ of the quitting strategy $\boldsymbol{b} = b$ for $b < \beta$. The value of the local-maximizing (in fact, ostensibly global-maximizing) value $b_{in} \approx 5.32$ is such that the confidence function $g(b)$ when using the quitting strategy $\boldsymbol{b} = (\infty, b_{in})$ is non-monotonic in the desired way: general increase with an intermediate period of decrease.

Appendix A Proofs

A.1 Proof of Proposition 1

Note that

$$g_{\rho_y}(b) = \frac{(1-y)\left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da}{y + (1-y)\left(\log\frac{1}{\eta}\right)\int_b^\infty \eta^a da} \tag{104}$$

$$= \frac{\left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da}{\frac{y}{1-y} + \left(\log\frac{1}{\eta}\right)\int_b^\infty \eta^a da}. \tag{105}$$

Recall that $\eta \in (0,1)$, and that $f(\cdot, b)$ is a continuous, non-negative function satisfying $f(b,b) = 1$. It follows that the integral

$$\mathfrak{w} = \left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da \tag{106}$$

is strictly positive, since we can find a positive-measure subset $[b, b+\varepsilon] \subset [b, \infty)$ on which the integrand

$$\left(\log\frac{1}{\eta}\right) f(a,b)\eta^a \tag{107}$$

is lower-bounded by a positive constant close to $f(b,b)\eta^b = \eta^b$. Also, the integral

$$\mathfrak{z} = \left(\log\frac{1}{\eta}\right)\int_b^\infty \eta^a da \tag{108}$$

is strictly positive.

Check that

$$\frac{\partial}{\partial y} g_{\rho_y}(b) = \frac{\partial}{\partial y}\frac{\mathfrak{w}}{\frac{y}{1-y} + \mathfrak{z}} = \frac{-\mathfrak{w}\frac{1}{(1-y)^2}}{\left(\frac{y}{1-y} + \mathfrak{z}\right)^2} = -\frac{\mathfrak{w}}{\left(y + \mathfrak{z}(1-y)\right)^2} < 0, \tag{109}$$

as desired. In particular, $g_{\rho_y}(b)$ is strictly monotonically decreasing in y, which yields the inequalities (45) as a corollary.

A.2 Proof of Proposition 2

To show part (a), we check that

$$\frac{d}{db}g_{\rho_0}(b) \tag{110}$$

is positive. First, we apply a change of variables to obtain

$$\begin{aligned}
g_{\rho_0}(b) &= \frac{\left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da}{\left(\log\frac{1}{\eta}\right)\int_b^\infty \eta^a da} \\
&= \frac{\left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da}{\eta^b} \\
&= \left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^{a-b} da \\
&= \left(\log\frac{1}{\eta}\right)\int_0^\infty f(b+m,b)\eta^m dm.
\end{aligned} \tag{111}$$

This equality can also be deduced from the memorylessness property of the exponential distribution ρ_0,

$$\rho_0(a) = \left(\log\frac{1}{\eta}\right)\eta^a. \tag{112}$$

Then, we differentiate the expression (111) with respect to b by Leibniz's integral rule, which yields

$$\begin{aligned}
\frac{d}{db}g_{\rho_0}(b) &= \frac{d}{db}\left(\left(\log\frac{1}{\eta}\right)\int_0^\infty f(b+m,b)\eta^m dm\right) \\
&= \left(\log\frac{1}{\eta}\right)\int_0^\infty \left(\frac{\partial}{\partial b}f(b+m,b)\right)\eta^m dm.
\end{aligned} \tag{113}$$

Recall that $\eta \in (0,1)$, and that $\frac{\partial}{\partial b}f(b+m,b) > 0$ by Assumption 1. Thus, the expression (113) is an integral of a positive and continuous function

$$\left(\log\frac{1}{\eta}\right)\left(\frac{\partial}{\partial b}f(b+m,b)\right)\eta^m \tag{114}$$

over $[0,\infty)$. Just as in Appendix A.1, we can find a positive-measure subset of $[0,\infty)$ on which the integrand is lower-bounded by a positive constant. Thus, the integral (113) is positive, as desired.

To show part (b), observe that

$$g_{\rho_y}(b) = \int_a f(a,b) d\rho_y^{\text{cond},a>b}(a), \tag{115}$$

where $\rho_y^{\text{cond},a>b}(a)$ denotes the conditional distribution of ρ_y conditional on $a > b$. Its p.d.f. is given by

$$\rho_y^{\text{cond},a>b}(a) = \frac{\rho_y(a)}{\int_{a>b} d\rho_y(a)} \quad \text{for} \quad a > b. \tag{116}$$

Observe that the conditional distribution $\rho_y^{\text{cond},a>b}$ places probability

$$\rho_y^{\text{cond},a>b}(\infty) = \frac{\rho_y^{\text{cond},a>b}(\infty)}{\int_{a>b} d\rho_y^{\text{cond},a>b}(a)} = \frac{y}{y + \left(\log \frac{1}{\eta}\right) \eta^b} \to 1 \tag{117}$$

on $a = \infty$ as $b \to \infty$. Equivalently, $\rho_y^{\text{cond},a>b}$ places probability converging to zero on the subset of finite difficulty values, (b, ∞), as $b \to \infty$. Since $f(\infty, b) = 0$, we can apply the dominated convergence theorem to conclude that

$$\begin{aligned} 0 \le \lim_{b\to\infty} g_{\rho_y}(b) &= \lim_{b\to\infty} \left(\int_{a\in(0,\infty)} f(a,b) d\rho_y^{\text{cond},a>b}(a) + 0 \cdot \rho_y^{\text{cond},a>b}(\infty) \right) \\ &\le \lim_{b\to\infty} \int_{a\in(0,\infty)} d\rho_y^{\text{cond},a>b}(a) \\ &= \int_{a\in(0,\infty)} \lim_{b\to\infty} d\rho_y^{\text{cond},a>b}(a) = \int_{a\in(0,\infty)} 0 \, da = 0, \end{aligned}$$

where we have set $\rho_y^{\text{cond},a>b}(a) = 0$ for $a \le b$. Thus, we have the desired equality

$$\lim_{b\to\infty} g_{\rho_y}(b) = 0. \tag{118}$$

To show part (c), we use the quotient rule and Leibniz's integral rule:

$$\begin{aligned} \frac{d}{db} g_{\rho_y}(b) &= \frac{d}{db} \frac{(1-y)\left(\log \frac{1}{\eta}\right) \int_b^\infty f(a,b) \eta^a da}{y + (1-y)\eta^b} \\ &= \frac{1}{\left(y + (1-y)\eta^b\right)^2} \end{aligned}$$

$$\cdot \Bigg(\left(y + (1-y)\eta^b\right)(1-y)\left(\log \frac{1}{\eta}\right)\left(-\eta^b + \int_b^\infty \frac{\partial}{\partial b} f(a,b)\eta^a da\right) \\ - (1-y)\left(\log\frac{1}{\eta}\right)\eta^b(1-y)\left(\log\frac{1}{\eta}\right)\int_b^\infty f(a,b)\eta^a da\Bigg). \tag{119}$$

Assumption 2 implies that that

$$-\eta^b + \int_b^\infty \frac{\partial}{\partial b} f(a,b)\eta^a da, \tag{120}$$

and thereby the entire expression (119) for $\frac{d}{db}g_{\rho_y}(b)$, is negative for all sufficiently large b, as desired.

A.3 Proof of Lemma 3

The proof of this lemma solely uses the fact that the unique solution

$$\begin{bmatrix} \hat{V}_{im}(b_{im}, b_{in}) \\ \hat{V}_{in}(b_{im}, b_{in}) \end{bmatrix} \tag{121}$$

to a nondegenerate system of equations

$$\begin{bmatrix} \mathfrak{a} & \mathfrak{b} \\ \mathfrak{c} & \mathfrak{d} \end{bmatrix} \begin{bmatrix} \hat{V}_{im}(b_{im}, b_{in}) \\ \hat{V}_{in}(b_{im}, b_{in}) \end{bmatrix} = \begin{bmatrix} \mathfrak{e} \\ \mathfrak{f} \end{bmatrix}. \tag{122}$$

is given by

$$\begin{bmatrix} \frac{\mathfrak{d}\mathfrak{e}-\mathfrak{b}\mathfrak{f}}{\mathfrak{a}\mathfrak{d}-\mathfrak{b}\mathfrak{c}} \\ \frac{\mathfrak{a}\mathfrak{f}-\mathfrak{c}\mathfrak{e}}{\mathfrak{a}\mathfrak{d}-\mathfrak{b}\mathfrak{c}} \end{bmatrix}. \tag{123}$$

Substituting the suitable expressions for the quantities $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}, \mathfrak{e}$, and $\mathfrak{f}$ completes our proof. Note that

$$\mathfrak{g} = \mathfrak{a}\mathfrak{d} - \mathfrak{b}\mathfrak{c}. \tag{124}$$

A.4 Proof of Proposition 4

Choose N large enough that the expected payoff deviation due to the procurement of alternative foraging opportunities in the model parametrization $\boldsymbol{M}(n)$ is less than $\varepsilon/3$ for all $n \geq N$. By possibly making N larger, the expected payoff deviation due to time-measurement experiments in the model parametrization $\boldsymbol{M}(n)$ is also less than $\varepsilon/3$.

Furthermore, by possibly making N even larger, the difference between the expected total payoff $V_n(\pi)$ in the model parametrization $\boldsymbol{M}(n)$—henceforward excluding deviations due to side opportunities and time-measurement costs—and that of its approximating continuous learning model $\boldsymbol{M}(\infty)$, given by $V_\infty(\boldsymbol{b}(\pi))$, is less than $\varepsilon/3$ for all $n \geq N$. To show this, we may as well assume that the task payoff of each learning period (say, the kth one) of the model parametrization $\boldsymbol{M}(n)$, given by

$$f(a(k), b(k)) \int_{T(k)}^{T(k+1)} \delta^t dt, \tag{125}$$

is obtained as a flow payoff of

$$\delta^t f(a(k), b(k)) dt. \tag{126}$$

We then define the function

$$\hat{V}_n(b_{im}, b_{in}) = q\hat{V}_{im,n} + (1-q)\hat{V}_{in,n} \tag{127}$$

in terms of the function $\hat{V}_{im,n}, \hat{V}_{in,n} : ((0,\infty) \cup \{\infty\})^2 \to [0,\infty)$, defined by

$$\hat{V}_{im,n}(b_{im}, b_{in}) = \frac{\mathfrak{d}_n \mathfrak{e}_n - \mathfrak{b}_n \mathfrak{f}_n}{\mathfrak{g}_n} \tag{128}$$

and

$$\hat{V}_{in,n}(b_{im}, b_{in}) = \frac{\mathfrak{a}_n \mathfrak{f}_n - \mathfrak{c}_n \mathfrak{e}_n}{\mathfrak{g}_n} \tag{129}$$

for

$$\mathfrak{a}_n = 1 - q\delta^{L^{-1}_{im,\infty}(b_{im})}\eta^{b_{im}} \tag{130}$$

$$\mathfrak{b}_n = -(1-q)\delta^{L^{-1}_{im,\infty}(b_{im})}\eta^{b_{im}}, \tag{131}$$

$$\mathfrak{c}_n = -q\delta^{L^{-1}_{in,\infty}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right), \tag{132}$$

$$\mathfrak{d}_n = 1 - (1-q)\delta^{L^{-1}_{in,\infty}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right), \tag{133}$$

$$\mathfrak{e}_n = \int_0^{b_{im}} \left(\int_0^{L_{im,\infty}^{-1}(a)} \delta^t f(a, L_{im,n}(t))dt + \int_{L_{im,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{im}(a) \\ + \int_{a > b_{im}} \left(\int_0^{L_{im,\infty}^{-1}(b_{im})} \delta^t f(a, L_{im,n}(t))dt \right) d\mu_{im}(a), \tag{134}$$

$$\mathfrak{f}_n = \int_0^{b_{in}} \left(\int_0^{L_{in,\infty}^{-1}(a)} \delta^t f(a, L_{in,n}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^{\infty} \delta^t dt \right) d\mu_{in}(a) \\ + \int_{a > b_{in}} \left(\int_0^{L_{in,\infty}^{-1}(b_{in})} \delta^t f(a, L_{in,n}(t))dt \right) d\mu_{in}(a), \tag{135}$$

and

$$\mathfrak{g}_n = 1 - \delta^{L_{in,\infty}^{-1}(b_{in})} \left(p + (1-p)\eta^{b_{in}}\right) + q \left(\delta^{L_{in,\infty}^{-1}(b_{in})} \left(p + (1-p)\eta^{b_{in}}\right) - \delta^{L_{im,\infty}^{-1}(b_{im})} \eta^{b_{im}} \right). \tag{136}$$

By construction, the functions $\hat{V}_n$ have the property that

$$V_n(\pi) = \hat{V}_n(b, b) \tag{137}$$

for a policy π represented by $\boldsymbol{b} = b$, and

$$V_n(\pi) = \hat{V}_n(b_{im}, b_{in}) \tag{138}$$

for a policy π represented by $\boldsymbol{b} = (b_{im}, b_{in})$.

Under the assumption that $\hat{V}_\infty$ and all functions $\hat{V}_n$ are continuous at $b = 0$, we will complete our proof. We have that $\{\hat{V}_n\}_{n \in \mathbb{N}}$ is a sequence of continuous functions on the compact space

$$\bar{\mathcal{Q}} \cup \{(0,0)\} = \{(b,b) : b \in [0, \beta]\} \cup \{(b_{im}, b_{in}) : b_{im}, b_{in} \in [\beta, \infty) \cup \{\infty\}\} \tag{139}$$

that is monotonically converging to $\hat{V}_\infty$, which is also continuous. Thus, this convergence is uniform by Dini's theorem. In particular, we have

$$\sup_{\pi \in \Pi} |V_n(\pi) - V_\infty(\boldsymbol{b}(\pi))| \le \sup_{(b_{im}, b_{in}) \in \bar{\mathcal{Q}} \cup \{(0,0)\}} \left| \hat{V}_n(b_{im}, b_{in}) - \hat{V}_n(b_{im}, b_{in}) \right| < \frac{\varepsilon}{3} \tag{140}$$

for sufficiently large n as desired, where we have used the fact that the set of all strategies $\boldsymbol{b}$ of the continuous learning model that represent policies $\pi \in \Pi$ of $\boldsymbol{M}(n)$ is a subset of $\bar{\mathcal{Q}}$. Our overall theorem statement then follows from the triangle inequality.

It remains to show that $\hat{V}_\infty$ (respectively, all functions $\hat{V}_n$), which are only defined for $b > 0$, can be continuously extended to $b = 0$. For this, it suffices to show that the constituent functions $\hat{V}_{im,\infty}$ and $\hat{V}_{in,\infty}$ (respectively, $\hat{V}_{im,n}$ and $\hat{V}_{im,n}$) can be continuously extended to $b = 0$. Observe that the numerator and denominator of each constituent function are both equal to zero at $b = 0$, which creates the *a priori* possible obstruction to continuity. However, by L'Hôspital's rule, if both the numerator and the denominator are differentiable at $b = 0$ and the derivative of the denominator has nonzero value at $b = 0$, then the limit of the function as $b \to 0$ is well-defined, as desired.

The derivative of the denominator $\mathfrak{g} = \mathfrak{g}_n$ at zero is computed by the product rule and chain rule:

$$\begin{aligned}\frac{d}{db}\mathfrak{g}(b,b)|_{b=0} &= (1-q)\left(\frac{\left(\log\frac{1}{\delta}\right)\delta^{L^{-1}_{in,\infty}(0)}}{\frac{d}{dt}L_{in,\infty}(0)}\left(p+(1-p)\eta^0\right)+\delta^{L^{-1}_{in,\infty}(0)}(1-p)\left(\log\frac{1}{\eta}\right)\eta^0\right)\\ &\quad+q\left(\frac{\left(\log\frac{1}{\delta}\right)\delta^{L^{-1}_{im,\infty}(0)}}{\frac{d}{dt}L_{in,\infty}(0)}\eta^0+\delta^{L^{-1}_{im,\infty}(0)}\left(\log\frac{1}{\eta}\right)\eta^0\right) \qquad (141)\\ &= (1-q)\left(\frac{\left(\log\frac{1}{\delta}\right)}{\frac{d}{dt}L_{in,\infty}(0)}\left(p+(1-p)\right)+(1-p)\left(\log\frac{1}{\eta}\right)\right)\\ &\quad+q\left(\frac{\left(\log\frac{1}{\delta}\right)}{\frac{d}{dt}L_{in,\infty}(0)}+\left(\log\frac{1}{\eta}\right)\right)>0. \qquad (142)\end{aligned}$$

To conclude via the product rule that the derivatives of the numerators of each of the functions $\hat{V}_{im,\infty}$ and $\hat{V}_{in,\infty}$ (respectively, $\hat{V}_{im,n}$, and $\hat{V}_{im,n}$) is well-defined at $b = 0$, it suffices to check whether the derivatives of $\mathfrak{e}$ (respectively, $\mathfrak{e}_n$) and $\mathfrak{f}$ (respectively, $\mathfrak{f}_n$) are well-defined at $b = 0$; this is because

$$\mathfrak{a} = \mathfrak{a}_n, \qquad (143)$$

$$\mathfrak{b} = \mathfrak{b}_n, \qquad (144)$$

$$\mathfrak{c} = \mathfrak{c}_n, \qquad (145)$$

and

$$\mathfrak{d} = \mathfrak{d}_n, \qquad (146)$$

are clearly differentiable via the chain rule. Indeed, Leibniz's integral rule yields that the derivatives of $\mathfrak{e}_n$ and $\mathfrak{f}_n$ are well-defined and given at $b = 0$ by

$$\begin{aligned} \frac{d}{db}\mathfrak{e}_n|_{b=0} &= \mu_{im}(0)\left(\int_0^{L^{-1}_{im,\infty}(0)} \delta^t f(a, L_{im,n}(t))dt + \int_{L^{-1}_{im,\infty}(0)}^{\infty} \delta^t dt\right) \\ &\qquad - \mu_{im}(0)\int_0^{L^{-1}_{im,\infty}(0)} \delta^t f(a, L_{im,n}(t))dt + \int_{a>0} \frac{\delta^{L^{-1}_{im,\infty}(0)} f(a,0)}{\frac{d}{dt}L_{im,\infty}(0)} d\mu_{im} \\ &= \left(\log\frac{1}{\eta}\right)\frac{1}{\log\frac{1}{\delta}} + \int_{a>0} \frac{\delta^{L^{-1}_{im,\infty}(0)} f(a,0)}{\frac{d}{dt}L_{im,\infty}(0)} d\mu_{im} \end{aligned} \tag{147}$$

and

$$\begin{aligned} \frac{d}{db}\mathfrak{f}_n|_{b=0} &= \mu_{in}(0)\left(\int_0^{L^{-1}_{in,\infty}(0)} \delta^t f(a, L_{in,n}(t))dt + \int_{L^{-1}_{in,\infty}(0)}^{\infty} \delta^t dt\right) \\ &\qquad - \mu_{in}(0)\int_0^{L^{-1}_{in,\infty}(0)} \delta^t f(a, L_{in,n}(t))dt + \int_{a>0} \frac{\delta^{L^{-1}_{in,\infty}(0)} f(a,0)}{\frac{d}{dt}L_{in,\infty}(0)} d\mu_{in} \\ &= (1-p)\left(\log\frac{1}{\eta}\right)\frac{1}{\log\frac{1}{\delta}} + \int_{a>0} \frac{\delta^{L^{-1}_{in,\infty}(0)} f(a,0)}{\frac{d}{dt}L_{in,\infty}(0)} d\mu_{in}. \end{aligned} \tag{148}$$

The calculations for $\mathfrak{e}$ and $\mathfrak{f}$ are analogous—the only difference being that the function $L_{j,n}(t)$ in the integrand is replaced with $L_{j,\infty}(t)$—and give the identical answers for the derivative at $b = 0$. The product rule thus yields the derivative of the numerators at $b = 0$, as needed.

A.5 Proof of Proposition 5

For every strategy $\boldsymbol{b} = (b_{im}, b_{in})$ such that $b_{im} < \infty$, we construct another strategy $\boldsymbol{b}'$ that achieves a strictly higher value $V_\infty(\boldsymbol{b}')$. This shows that a necessary condition for $\boldsymbol{b} = (b_{im}, b_{in})$ to maximize V_∞ is that $b_{im} = \infty$. Note that the constructed strategy $\boldsymbol{b}'$ will not be of the form $\boldsymbol{b}' = (b'_{im}, b'_{in})$, i.e., will not repeat the same quitting strategy for every drawn task.

Consider the probability measure μ^∞ on the sample space of sequences of tasks drawn i.i.d. from μ (some of which may not be drawn if the student quits finitely many times),

$$\Omega = \mathcal{U}^\infty. \tag{149}$$

The distribution is defined as follows. Let $\mathcal{F}$ denote the σ-algebra generated by the algebra

$$\mathcal{F}_0 = \bigcup_{n=1}^{\infty} \mathcal{F}_n, \tag{150}$$

where $\mathcal{F}_n$ denotes the collection of events whose occurrence can be determined by the results of the first n draws. The probability distribution μ on $\mathcal{U}$ canonically endows $\mathcal{F}$ with a probability measure μ^{∞}, which is used to defined the compute the expected value of the payoff.

Let $V_{\infty}(\boldsymbol{b}'', \omega)$ denote the total payoff when the student uses a strategy $\boldsymbol{b}''$ and the sequence of task types is $\omega \in \Omega$. Then, the total payoff $V_{\infty}(\boldsymbol{b}'')$ is given by

$$V_{\infty}(\boldsymbol{b}'') = \int_{\Omega} V_{\infty}(\boldsymbol{b}'') d\mu^{\infty}(\omega). \tag{151}$$

We modify $\boldsymbol{b} = (b_{im}, b_{in})$ to obtain the alternative strategy

$$\boldsymbol{b}' = (([q, b'_{im}; (1-q)p, b'_{in}], b_{in}), (b_{im}, b_{in}), (b_{im}, b_{in}), \ldots), \tag{152}$$

where the first factor

$$[q, b'_{im}; (1-q), b'_{in}] \tag{153}$$

denote the probabilistic quitting strategy of, assuming learning has not completed by then, quitting with probability q at

$$b'_{im} = L_{im,\infty}\left(2L^{-1}_{im,\infty}(b_{im})\right) \tag{154}$$

and quitting with probability $(1-q)$ at

$$b'_{in} = L_{im,\infty}\left(L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{im,\infty}(b_{in})\right). \tag{155}$$

The probabilistic strategy $\boldsymbol{b}'$ can be written as a combination of two deterministic strategies:

$$\boldsymbol{b}'_{im} = ((b'_{im}, b_{in}), (b_{im}, b_{in}), (b_{im}, b_{in}, \ldots) \tag{156}$$

with probability q and

$$\boldsymbol{b}'_{im} = ((b'_{in}, b_{in}), (b_{im}, b_{in}), (b_{im}, b_{in}, \ldots) \tag{157}$$

with probability $1 - q$.

We will show that

$$V_\infty(\boldsymbol{b}) = \int_\Omega V_\infty(\boldsymbol{b}) d\mu^\infty(\omega) \tag{158}$$

is strictly less than

$$V_\infty(\boldsymbol{b}') = \int_\Omega V_\infty(\boldsymbol{b}') d\mu^\infty(\omega), \tag{159}$$

thus showing that $\boldsymbol{b}'$ strictly outperforms $\boldsymbol{b}$.

First, we partition the sample space Ω into subsets

$$\Omega = \Omega_1 \cup \Omega_2, \tag{160}$$

defined by

$$\Omega_1 = \{\omega = ((j_1, a_1), \ldots) : j_1 = in \text{ or } a_1 \leq b_{im}\} \tag{161}$$

$$\Omega_2 = \{\omega = ((j_1, a_1), \ldots) : j_1 = im \text{ and } a_1 > b_{im}\}. \tag{162}$$

Note that

$$\int_{\Omega_1} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty(\omega) = \int_{\Omega_1} V_\infty(\boldsymbol{b}', \omega) d\mu^\infty(\omega). \tag{163}$$

Indeed, if $j_1 = im$ and $a_1 \leq b_{im}$ for $\omega \in \Omega_1$, then both $\boldsymbol{b}$ and $\boldsymbol{b}'$ learn the first task until completion and stick with it forever; and if $j_1 = in$, the strategies $\boldsymbol{b}$ and $\boldsymbol{b}'$ play in the same way for such a task sequence ω.

It thus suffices to show that

$$\int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty < \int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}', \omega) d\mu^\infty. \tag{164}$$

Partition Ω_2 into subsets

$$\Omega_2 = \Omega_3 \cup \Omega_4 \cup \Omega_5 \tag{165}$$

defined by

$$\Omega_3 = \{\omega = ((j_1, a_1), (j_2, a_2), \ldots) : j_1 = im, a_1 > b_{im}, \text{ and } j_2 = im\} \tag{166}$$

$$\Omega_4 = \{\omega = ((j_1, a_1), (j_2, a_2), \ldots) : j_1 = im, a_1 > b_{im}, j_2 = in, \text{ and } a_2 < \infty\} \tag{167}$$

and

$$\Omega_5 = \{\omega = ((j_1, a_1), (j_2, a_2), \ldots) : j_2 = in, a_2 = \infty, j_2 = in, \text{ and } a_2 = \infty\}. \tag{168}$$

It suffices to show that

$$\int_{\omega \in \Omega_3} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty < q \int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}'_{im}, \omega) d\mu^\infty, \tag{169}$$

$$\int_{\omega \in \Omega_4} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty < (1-q)(1-p) \int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}'_{in}, \omega) d\mu^\infty, \tag{170}$$

and

$$\int_{\omega \in \Omega_5} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty < (1-q)p \int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}'_{in}, \omega) d\mu^\infty, \tag{171}$$

since $\boldsymbol{b}'$ plays as the strategy $\boldsymbol{b}'_{im}$ with probability q (the proportion of Ω_3 in Ω_2) and as the strategy $\boldsymbol{b}'_{in}$ with probability (the proportion of Ω_4 and Ω_5 combined in Ω_2).

We first show inequality (169). Check that the left-hand side is given by

$$\begin{aligned} \int_{\omega \in \Omega_3} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty &= \int_{\bar{\omega} \in \Omega'_3} V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), \bar{\omega}))) \left(\int_{\{(im, a_1) : a_1 > b_{im}\}} d\mu \right) d\mu^\infty \\ &= q\eta^{b_{im}} \int_{\bar{\omega} \in \Omega'_3} V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), \bar{\omega})) d\mu^\infty, \end{aligned} \tag{172}$$

where $\bar{\omega} \in \Omega'_3$ parametrizes the task subsequence of $\omega \in \Omega_3$ given by

$$\bar{\omega} = ((j_2, a_2), (j_3, a_3), \ldots), \tag{173}$$

$b_{im} + \varepsilon$ is an arbitrary task difficulty level greater than b_{im},

$$V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), \bar{\omega}))) \tag{174}$$

does not depend on the choice of $b_{im} + \varepsilon$, and we have an isomorphism of probability spaces

$$\Omega'_3 \cong \{(im, a) : a > 0\} \times \mathcal{U}^\infty. \tag{175}$$

Next, check that the right-hand side can be written as

$$\begin{aligned} q\int_{\omega\in\Omega_2} V_\infty(\boldsymbol{b}'_{im},\omega)d\mu^\infty &= q\int_{\{(im,a_1):a_1>b_{im}\}}\left(\int_{\hat{\omega}\in\Omega_2'} V_\infty\left(\boldsymbol{b}',((j_1,a_1),\hat{\omega})\right)d\mu^\infty\right)d\mu \\ &= q\eta^{b_{im}}\int_{\{(im,a):a>0\}}\left(\int_{\hat{\omega}\in\Omega_2'} V_\infty\left(\boldsymbol{b}',((j_1,a+b_{im}),\hat{\omega})\right)d\mu^\infty\right)d\mu, \end{aligned} \tag{176}$$

where $\hat{\omega}\in\Omega_2'$ parametrizes the task subsequence of $\omega\in\Omega_2$ given by

$$\hat{\omega}=((j_2,a_2),(j_3,a_3),\ldots), \tag{177}$$

and we have an isomorphism of probability spaces

$$\Omega_2'\cong\mathcal{U}^\infty. \tag{178}$$

Using the isomorphisms, we reduce our inequality (169) to the following:

$$\begin{aligned} &\int_{((j_2,a),(j_3,a_3),\ldots)\in\{(im,a')\ :a'>0\}\times\mathcal{U}^\infty} V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(j_2,a),(j_3,a_3)\ldots))d\mu^\infty \\ &\qquad<\int_{((j_1,a),(j_2,a_2),\ldots)\in\{(im,a')\ :a'>0\}\times\mathcal{U}^\infty} V_\infty(\boldsymbol{b}',((j_1,a+b_{im}),(j_2,a_2)\ldots))d\mu^\infty. \end{aligned} \tag{179}$$

There is a clear isomorphism of the probability space of task sequences

$$((j_2,a),(j_3,a_3),\ldots)\in\{(im,a')\ :a'>0\}\times\mathcal{U}^\infty \tag{180}$$

and the probability space

$$((j_1,a),(j_2,a_2),\ldots)\in\{(im,a')\ :a'>0\}\times\mathcal{U}^\infty. \tag{181}$$

It suffices to show that the strict inequality holds for the one-to-one-corresponding integrands in this isomorphism, which we will refer to as the left-hand-side value function

$$V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(j_2,a),(j_3,a_3)\ldots)) \tag{182}$$

and the right-hand-side value function

$$V_\infty(\boldsymbol{b'}, ((j_1, a + b_{im}), (j_2, a_2)\ldots)) \tag{183}$$

We need to show that

$$V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), (im, a), (j'_3, a'_3), \ldots)) < V_\infty(\boldsymbol{b'}, ((im, a + b_{im}), (j'_2, a'_2), (j'_3, a'_3), \ldots)) \tag{184}$$

Note that the sub-payoff values in the subinterval of time

$$[0, L^{-1}_{im,\infty}(b_{im})) \tag{185}$$

for both value functions are identical. This is because the first task is of type $j = im$ and is learned to the point of time $L^{-1}_{im,\infty}(b_{im})$ for both value functions.

Also, conditional on the assumption that the task that is learned at time $t = L^{-1}_{im,\infty}(b_{im})$ (second task and first task, respectively) does not learn to completion—that $a_1 < b_{im}$ and $a_1 + b_{im} < b'_{im}$, respectively—the sub-payoff values in the subinterval of time

$$[2L^{-1}_{im,\infty}(b_{im}), \infty) \tag{186}$$

are also identical for both value functions. This is because conditional on this assumption, the aforementioned task is quit at time $t = 2L^{-1}_{im,\infty}(b_{im})$, after which the payoff in the remaining time is the same.

Next, we show that if the task that is learned at time $t = L^{-1}_{im,\infty}(b_{im})$ learns to completion for the left-hand-side value function in that $a_1 < b_{im}$, then it also learns to completion for the right hand-side value function in that $a_1 + b_{im} < b'_{im}$. This is a consequence of the assumption that $L_{im,\infty}(t)$ is convex. It follows that at time $t = L^{-1}_{im,\infty}(b_{im})$, the difference a in knowledge that is required to complete the task learning requires less (or equal) time for the right-hand-side value function, spanning

$$t = L^{-1}_{im,\infty}(b_{im}) \quad \text{to} \quad t = L^{-1}_{im,\infty}(b_{im} + a); \tag{187}$$

than the time required to complete the task learning for the left-hand-side value function, spanning

$$t = L^{-1}_{im,\infty}(b_{im}) \quad \text{to} \quad t = L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{im,\infty}(a). \tag{188}$$

Indeed, our assumption that $L_{im,\infty}(t)$ is convex yields the fact that $L_{im,\infty}^{-1}(b)$ is concave, which yields

$$L_{im,\infty}^{-1}(b_{im}+a) \leq L_{im,\infty}^{-1}(b_{im}) + L_{im,\infty}^{-1}(a). \tag{189}$$

If learning of this task completes for the left-hand-side, then it also completes for the right-hand-side; consequently, no future tasks are drawn, and the sub-payoff values for the subperiod of time (186) are equal. On the other hand, if learning of this task completes for the right-hand-side value function, then no future tasks are drawn for it (but may be drawn for the left-hand-side value function); consequently, the sub-payoff values for the subperiod of time (186) automatically satisfy the desired direction of inequality.

Moreover, the respective sub-payoff values in the remaining subperiod of time

$$[L_{im,\infty}^{-1}(b_{im}), 2L_{im,\infty}^{-1}(b_{im})) \tag{190}$$

are given by

$$\int_{L_{im,\infty}^{-1}(b_{im})}^{2L_{im,\infty}^{-1}(b_{im})} \delta^t \left\{ \begin{array}{ll} f(a, L_{im,\infty}(t - L_{im,\infty}^{-1}(b_{im}))) & \text{for } t < L_{im,\infty}^{-1}(b_{im}) + L_{im,\infty}^{-1}(a) \\ 1 & \text{for } t \geq L_{im,\infty}^{-1}(b_{im}) + L_{im,\infty}^{-1}(a) \end{array} \right\} dt \tag{191}$$

for the left-hand-side value function and

$$\int_{L_{im,\infty}^{-1}(b_{im})}^{2L_{im,\infty}^{-1}(b_{im})} \delta^t \left\{ \begin{array}{ll} f(b_{im}+a, L_{im,\infty}(t)) & \text{for } t < L_{im,\infty}^{-1}(b_{im}+a) \\ 1 & \text{for } t \geq L_{im,\infty}^{-1}(b_{im}+a) \end{array} \right\} dt \tag{192}$$

for the right-hand-side value function. It follows from the inequalities (189),

$$\begin{aligned} f(a, L_{im,\infty}(t - L_{im,\infty}^{-1}(b_{im}))) &< f(b_{im}+a, b_{im} + L_{im,\infty}(t - L_{im,\infty}^{-1}(b_{im}))) \\ &\leq f(b_{im}+a, L_{im,\infty}(t)), \end{aligned} \tag{193}$$

and

$$f(a,b) \leq 1 \tag{194}$$

that the sub-payoff value (191) of the left-hand-side value function is strictly less than that (192) of the right-hand-side value function.

We have overall shown the inequality of integrands (184), which implies the inequality

(176), and thereby, the inequality (169).

The second of our desired inequality (170) will be shown analogously. Check that the left-hand side is given by

$$
\begin{aligned}
&\int_{\omega\in\Omega_4} V_\infty(\boldsymbol{b},\omega)d\mu^\infty \\
&= \int_{((in,a_2),\bar{\omega})\in\{(in,a_2):a_2\in(0,\infty)\}\times\Omega_4'} V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a_2),\bar{\omega}))\left(\int_{\{(im,a_1):a_1>b_{im}} d\mu\right) d\mu^\infty \\
&= q\eta^{b_{im}} \int_{((in,a_2),\bar{\omega})\in\{(in,a_2):a_2\in(0,\infty)\}\times\Omega_4'} V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a_2),\bar{\omega}))d\mu^\infty \\
&= q\eta^{b_{im}}(1-q)(1-p)\int_{a_2\in(0,\infty)}\left(\log\frac{1}{\eta}\right)\eta^{a_2}\left(\int_{\bar{\omega}\in\Omega_4'} V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a_2),\bar{\omega}))d\mu^\infty\right)da_2,
\end{aligned}
\tag{195}
$$

where $\bar{\omega}\in\Omega_4'$ parametrizes the task subsequence of $\omega\in\Omega_4$,

$$
\bar{\omega} = ((j_3,a_3),(j_4,a_4),\ldots),
\tag{196}
$$

$b_{im}+\varepsilon$ is an arbitrary task difficulty level greater than b_{im},

$$
V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a_2),\bar{\omega}))
\tag{197}
$$

does not depend on the choice of $b_{im}+\varepsilon$, and we have an isomorphism of probability spaces

$$
\Omega_4' \cong \mathcal{U}^\infty.
\tag{198}
$$

Next, check that the right-hand-side inequality can be written as

$$
\begin{aligned}
&(1-q)(1-p)\int_{\omega\in\Omega_2} V_\infty(\boldsymbol{b}_{in}',\omega)d\mu^\infty \\
&\qquad= (1-q)(1-p)\int_{\{(im,a_1):a_1>b_{im}\}}\left(\int_{\hat{\omega}\in\Omega_2'} V_\infty(\boldsymbol{b}_{in}',((im,a_1),\hat{\omega}))d\mu^\infty\right)d\mu \\
&\qquad= (1-q)(1-p)q\eta^{b_{im}}\int_{a\in(0,\infty)}\left(\log\frac{1}{\eta}\right)\eta^{a}\left(\int_{\hat{\omega}\in\Omega_2'} V_\infty(\boldsymbol{b}_{in}',((im,a+b_{im}),\hat{\omega}))d\mu^\infty\right)da
\end{aligned}
\tag{199}
$$

Using the isomorphisms (219) and (175), we reduce our inequality (170) to

$$\int_{(a,(j_3,a_3),\ldots)\in(0,\infty)\times\mathcal{U}^\infty} V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a),(j_3,a_3)\ldots))d\mu^\infty d\mu_\eta \\ < \int_{(a,(j_2,a_2),\ldots)\in(0,\infty)\times\mathcal{U}^\infty} V_\infty(\boldsymbol{b}',((im,a+b_{im}),(j_2,a_2)\ldots))d\mu^\infty d\mu_\eta, \tag{200}$$

where $\mu_\eta = \mu_{im} = \mu_{in}|_{a<\infty}$ denotes the exponential distribution of decay factor η on $(0,\infty)$.

There is a clear isomorphism of the probability space of task sequences

$$(a,(j_3,a_3),\ldots) \in (0,\infty)\times\mathcal{U}^\infty \tag{201}$$

and the probability space

$$(a,(j_2,a_2),\ldots) \in (0,\infty)\times\mathcal{U}^\infty. \tag{202}$$

It suffices to show that the strict inequality holds for the one-to-one-corresponding integrands in this isomorphism, which we will refer to as the left-hand-side value function

$$V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a),(j_3,a_3)\ldots)) \tag{203}$$

and the right-hand-side value function

$$V_\infty(\boldsymbol{b}',((im,a+b_{im}),(j_2,a_2)\ldots)). \tag{204}$$

We need to show that

$$V_\infty(\boldsymbol{b},((im,b_{im}+\varepsilon),(in,a),(j_3,a_3)\ldots)) < V_\infty(\boldsymbol{b}',((im,a+b_{im}),(j_2,a_2)\ldots)). \tag{205}$$

Just as before, the sub-payoff-values in the subinterval of time

$$[0,L^{-1}_{im,\infty}(b_{im})) \tag{206}$$

for both value functions are identical.

Also, similarly to before, conditional on the assumption that task that is learned at time $t = L^{-1}_{im,\infty}(b_{im})$ (second task and first task, respectively) does not learn to completion—that

$a_1 < b_{in}$ and $a_1 + b_{im} < b'_{in}$, respectively—the sub-payoff values in the subinterval of time

$$[L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(b_{in}), \infty) \tag{207}$$

are identical for both value functions.

Next, we show that if the task that is learned at time $t = L^{-1}_{im,\infty}(\infty)$ learns to completion for the left-hand sidevaue function in that $a_1 < b_{in}$, then it also learns to completion for the right-hand-side value function in that $a_1 + b_{im} < b'_{in}$. This is a consequence of two assumptions: the assumption that $L_{im,\infty}(t)$ is convex (equivalently, that $L^{-1}_{im,\infty}(b)$ is concave) and the assumption that $L_{in,\infty}(t) \leq L_{im,\infty}(t)$ (equivalently, that $L^{-1}_{im,\infty}(b) \leq L^{-1}_{in,\infty}(b)$). It follows that at time $t = L^{-1}_{im,\infty}(b_{im})$, the difference a in knowledge that is required to complete the task learning requires less (or equal) time for the right-hand-side value function, spanning

$$t = L^{-1}_{im,\infty}(b_{im}) \quad \text{to} \quad t = L^{-1}_{im,\infty}(b_{im} + a), \tag{208}$$

than the time required to complete the task learning for the left-hand-side value function, spanning

$$t = L^{-1}_{im,\infty}(b_{im}) \quad \text{to} \quad t = L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(a), \tag{209}$$

Indeed, our two aforementioned assumptions together yield

$$L^{-1}_{im,\infty}(b_{im} + a) \leq L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{im,\infty}(a) \leq L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(a). \tag{210}$$

If learning of this task completes for the left-hand-side, then it also completes for the right-hand-side; consequently, no future tasks are drawn, and the sub-payoff values for the subperiod of time (207) are equal. On the other hand, if learning of this task completes for the right-hand-side value function, then no future tasks are draw for it (but may be drawn for the left-hand-side value function); consequently, the sub-payoff values for the subperiod of time (207) automatically satisfy the desired direction of inequality.

Finally, the respective sub-payoff values in the remaining subperiod of time

$$[L^{-1}_{im,\infty}(b_{im}), L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(b_{in})) \tag{211}$$

are given by

$$\int_{L^{-1}_{im,\infty}(b_{im})}^{L^{-1}_{im,\infty}(b_{im})+L^{-1}_{in,\infty}(b_{in})} \delta^t \left\{ \begin{array}{ll} f(a, L_{in,\infty}(t - L^{-1}_{im,\infty}(b_{im}))) & \text{for } t < L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(a) \\ 1 & \text{for } t \geq L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(a) \end{array} \right\} dt \tag{212}$$

for the left-hand-side value function and

$$\int_{L^{-1}_{im,\infty}(b_{im})}^{L^{-1}_{im,\infty}(b_{im})+L^{-1}_{in,\infty}(b_{in})} \delta^t \left\{ \begin{array}{ll} f(b_{im}+a, L_{im,\infty}(t)) & \text{for } t < L^{-1}_{im,\infty}(b_{im}+a) \\ 1 & \text{for } t \geq L^{-1}_{im,\infty}(b_{im}+a) \end{array} \right\} dt \tag{213}$$

for the right-hand-side value function. It follows from the inequalities (210),

$$\begin{aligned} f(a, L_{in,\infty}(t - L^{-1}_{im,\infty}(b_{im}))) &\leq f(a, L_{im,\infty}(t - L^{-1}_{im,\infty}(b_{im}))) \\ &< f(b_{im}+a, b_{im} + L_{im,\infty}(t - L^{-1}_{im,\infty}(b_{im}))) \\ &\leq f(a, L_{im,\infty}(t)), \end{aligned} \tag{214}$$

and $f(a,b) \leq 1$ that the sub-payoff value (212) of the left-hand-side value function is strictly less than that (213) of the right-hand-side value function.

Overall, we have shown the inequality of integrands (205), which implies the inequality (200) and thereby, the inequality (170).

It remains to show the inequality (171). Check that the left-hand side is given by

$$\begin{aligned} &\int_{\omega\in\Omega_5} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty \\ &= \int_{((in,a_2),\bar{\omega})\in\{(in,\infty)\}\times\Omega_5'} V_\infty(\boldsymbol{b}, ((im, b_{im}+\varepsilon), (in, a_2), \bar{\omega}))) \left(\int_{\{(im,a_1):a_1>b_{im}} d\mu \right) d\mu^\infty \\ &= q\eta^{b_{im}} \int_{((in,a_2),\bar{\omega})\in\{(in,\infty)\}\times\Omega_5'} V_\infty(\boldsymbol{b}, ((im, b_{im}+\varepsilon), (in, a_2), \bar{\omega})) d\mu^\infty \\ &= q\eta^{b_{im}}(1-q)p \left(\int_{\bar{\omega}\in\Omega_5'} V_\infty(\boldsymbol{b}, ((im, b_{im}+\varepsilon), (in, \infty), \bar{\omega})) d\mu^\infty \right) &&(215) \\ &= q\eta^{b_{im}}(1-q)p \int_{\bar{a}\in(0,\infty)} \left(\int_{\bar{\omega}\in\Omega_5'} V_\infty(\boldsymbol{b}, ((im, b_{im}+\varepsilon), (in, \infty), \bar{\omega})) d\mu^\infty \right) d\mu_\eta, &&(216) \end{aligned}$$

where $\bar{\omega} \in \Omega_5'$ parametrizes the task subsequence of $\omega \in \Omega_5$,

$$\bar{\omega} = ((j_3, a_3), (j_4, a_4), \ldots), \tag{217}$$

$b_{im} + \varepsilon$ is an arbitrary task difficulty level greater than b_{im},

$$V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), (in, \infty), \bar{\omega})) \tag{218}$$

does not depend on the choice of $b_{im} + \varepsilon$, we have an isomorphism of probability spaces

$$\Omega_5' \cong \mathcal{U}^\infty, \tag{219}$$

and $\bar{a}$, distributed as μ_η, is a dummy variable. Next, check that the right-hand side of the inequality

$$\begin{aligned} &(1-q)p \int_{\omega \in \Omega_2} V_\infty(\boldsymbol{b}_{in}', \omega) d\mu^\infty \\ &\qquad = (1-q)p \int_{\{(im, a_1): a_1 > b_{im}\}} \left(\int_{\hat{\omega} \in \Omega_2'} V_\infty(\boldsymbol{b}_{in}', ((im, a_1), \hat{\omega})) d\mu^\infty \right) d\mu \\ &\qquad = (1-q)pq\eta^{b_{im}} \int_{a \in (0,\infty)} \left(\int_{\hat{\omega} \in \Omega_2'} V_\infty(\boldsymbol{b}_{in}', ((im, a + b_{im}), \hat{\omega})) d\mu^\infty \right) d\mu_\eta. \end{aligned} \tag{220}$$

There is a clear isomorphism of the probability space of task sequences

$$(a, (j_3, a_3), \ldots) \in (0, \infty) \times \mathcal{U}^\infty \tag{221}$$

and the probability space

$$(a, (j_2, a_2), \ldots) \in (0, \infty) \times \mathcal{U}^\infty. \tag{222}$$

It suffices to show that the strict inequality holds for the one-to-one-corresponding integrands in this isomorphism, which we will refer to as the left-hand-side value function

$$V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), (in, \infty), (j_3, a_3) \ldots)) \tag{223}$$

and the right-hand-side value function

$$V_\infty(\boldsymbol{b}', ((im, a + b_{im}), (j_2, a_2) \ldots)). \tag{224}$$

We need to show that

$$V_\infty(\boldsymbol{b}, ((im, b_{im} + \varepsilon), (in, \infty), (j_3, a_3) \ldots)) < V_\infty(\boldsymbol{b}', ((im, a + b_{im}), (j_2, a_2) \ldots)). \tag{225}$$

Just as before, the sub-payoff-values in the subinterval of time

$$[0, L^{-1}_{im,\infty}(b_{im})) \tag{226}$$

for both value functions are identical.

Also, similarly to before, conditional on the assumption that task that is learned at time $t = L^{-1}_{im,\infty}(b_{im})$ (second task and first task, respectively) does not learn to completion—that $a_1 < b_{in}$ and $a_1 + b_{im} < b'_{in}$, respectively—the sub-payoff values in the subinterval of time

$$[L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(b_{in}), \infty) \tag{227}$$

are identical for both value functions.

Next, we show that if the task that is learned at time $t = L^{-1}_{im,\infty}(\infty)$ learns to completion for the left-hand sidevaue function in that $a_1 < b_{in}$, then it also learns to completion for the right-hand-side value function in that $a_1 + b_{im} < b'_{in}$. In fact, note that learning for this task can never complete for the left-hand-side value function, since the task difficulty is $a = \infty$. Thus, this step is trivially satisfied. consequently, the sub-payoff values for the subperiod of time (227) automatically satisfy the desired direction of inequality.

Finally, the respective sub-payoff values in the remaining subperiod of time

$$[L^{-1}_{im,\infty}(b_{im}), L^{-1}_{im,\infty}(b_{im}) + L^{-1}_{in,\infty}(b_{in}),] \tag{228}$$

are given by

$$\int_{L^{-1}_{im,\infty}(b_{im})}^{L^{-1}_{im,\infty}(b_{im})+L^{-1}_{in,\infty}(b_{in})} \delta^t f(\infty, L_{in,\infty}(t - L^{-1}_{im,\infty}(b_{im})))dt = 0 \tag{229}$$

for the left-hand-side value function and

$$\int_{L_{im,\infty}^{-1}(b_{im})}^{L_{im,\infty}^{-1}(b_{im})+L_{in,\infty}^{-1}(b_{in})} \delta^t \left\{ \begin{array}{ll} f(b_{im}+a, L_{im,\infty}(t)) & \text{for } t < L_{im,\infty}^{-1}(b_{im}+a) \\ 1 & \text{for } t \geq L_{im,\infty}^{-1}(b_{im}+a) \end{array} \right\} dt \tag{230}$$

for the right-hand-side value function. It follows that the sub-payoff value (229) of the left-hand-side value function is strictly less than that (230) of the right-hand-side value function.

We have overall shown (225). This shows the inequality (220), and thereby, the desired inequality (171). This completes our proof.

A.6 Proof of Proposition 6

We use a similar proof strategy as that of the proof of Proposition 5. For every strategy $\boldsymbol{b} = (b_{im}, \infty)$, we construct another strategy $\boldsymbol{b}'$, not of the form $\boldsymbol{b}' = (b'_{im}, b'_{in})$, that achieves a strictly higher value $V_\infty(\boldsymbol{b}')$. This shows that a necessary condition for $\boldsymbol{b} = (b_{im}, b_{in})$ to maximize V_∞ is that $b_{in} = \infty$.

The modification $\boldsymbol{b}'$ is defined by

$$\boldsymbol{b}' = ((b_{im}, b'), (b_{im}, \infty), (b_{im}, \infty), \ldots)\,, \tag{231}$$

for a value b' that will be specified later.

We partition the sample space Ω into subsets

$$\Omega = \Omega_6 \cup \Omega_7 \tag{232}$$

defined by

$$\Omega_6 = \{\omega = ((j_1, a_1), \ldots) : j_1 = im \text{ or } a_1 \leq b'\} \tag{233}$$

and

$$\Omega_7 = \{\omega = ((j_1, a_1), \ldots) : j_1 = in \text{ and } a_1 > b'\}. \tag{234}$$

Note that

$$\int_{\Omega_6} V_\infty(\boldsymbol{b}, \omega) d\mu^\infty = \int_{\Omega_6} V_\infty(\boldsymbol{b}', \omega) d\mu^\infty \tag{235}$$

Indeed, if $j_1 = in$ and $a_1 \leq b'$ for $\omega \in \Omega_1$, then both $\boldsymbol{b}$ and $\boldsymbol{b}'$ learn the first task until completion and stick with it forever; and if $j_1 = im$, the strategies $\boldsymbol{b}$ and $\boldsymbol{b}'$ play in the same way for such a task sequence ω.

It thus suffices to show that

$$\int_{\omega\in\Omega_7} V_\infty(\boldsymbol{b},\omega)d\mu^\infty < \int_{\omega\in\Omega_7} V_\infty(\boldsymbol{b}',\omega)d\mu^\infty. \tag{236}$$

Observe that for each $\omega \in \Omega_7$, the sub-payoff value of the value function $V_\infty(\boldsymbol{b},\omega)$ and that of the value function $V_\infty(\boldsymbol{b}',\omega)$ for the subperiod of time

$$[0, L^{-1}_{in,\infty}(b')) \tag{237}$$

are identical, since both strategies learn the first task during this subperiod.

The key insight is that the sub-payoff value of the value function $V_\infty(\boldsymbol{b}',\omega)$ in the remaining subperiod of time

$$[L^{-1}_{in,\infty}(b'), \infty) \tag{238}$$

is always given by

$$\delta^{L^{-1}_{in,\infty}(b')} V_\infty(\boldsymbol{b}'',\bar{\omega}), \tag{239}$$

where

$$\boldsymbol{b}'' = ((b_{im},\infty),(b_{im},\infty),\ldots) \tag{240}$$

and

$$\bar{\omega} = ((j_2,a_2),\ldots) \tag{241}$$

are obtained from $\boldsymbol{b}'$ and ω, respectively, by truncating the leftmost term. It follows that the integral of this sub-payoff value over Ω_7 is

$$\begin{aligned}\int_{\omega\in\Omega_7} \delta^{L^{-1}_{in,\infty}(b')} V_\infty(\boldsymbol{b}'',\bar{\omega})d\mu^\infty &= \delta^{L^{-1}_{in,\infty}(b')} \int_{\{(in,a_1):a_1>b'\}} \int_{\bar{\omega}\in\mathcal{U}^\infty} V_\infty(\boldsymbol{b}'',\bar{\omega})d\mu^\infty d\mu \\ &= \delta^{L^{-1}_{in,\infty}(b')} \left(p + (1-p)\eta^{b'}\right) V_\infty(\boldsymbol{b}''). \end{aligned} \tag{242}$$

In contrast, the integral of the sub-payoff value of the value function $V_\infty(\boldsymbol{b},\omega)$ in the subperiod (238) is given by

$$\int_{\omega\in\Omega_7} \left(\int_{L^{-1}_{in,\infty}(b')}^{\infty} f(a_1, L_{in,\infty}(t))dt \right) d\mu^\infty$$

$$\begin{aligned}
&= \int_{\{\omega\in\Omega_7:a_1=\infty\}} \left(\int_{L_{in,\infty}^{-1}(b')}^{\infty} f(a_1, L_{in,\infty}(t))dt\right) d\mu^{\infty} \\
&\qquad + \int_{\{\omega\in\Omega_7:a_1<\infty\}} \left(\int_{L_{in,\infty}^{-1}(b')}^{\infty} f(a_1, L_{in,\infty}(t))dt\right) d\mu^{\infty} \\
&= \int_{\{\omega\in\Omega_7:a_1<\infty\}} \left(\int_{L_{in,\infty}^{-1}(b')}^{\infty} f(a_1, L_{in,\infty}(t))dt\right) d\mu^{\infty} \\
&\le \int_{\{\omega\in\Omega_7:a_1<\infty\}} d\mu^{\infty} \left(\int_{L_{in,\infty}^{-1}(b')}^{\infty} 1dt\right) \\
&= \left((1-p)\eta^{b'}\right) \delta^{L_{in,\infty}^{-1}(b')} \frac{1}{\log\frac{1}{\delta}}. \qquad (243)
\end{aligned}$$

Note that as $b' \to \infty$, the expression (242) divided by $\delta^{L_{in,\infty}^{-1}(b')}$ converges to

$$pV_{\infty}(\boldsymbol{b}'') > 0, \qquad (244)$$

whereas the upper bound (243) divided by $\delta^{L_{in,\infty}^{-1}(b')}$ converges to 0. This shows that for b' sufficiently large, the inequality (236) holds.

For $\boldsymbol{b} = b = \infty$, since $V_{\infty}(b) = V_{\infty}(b,b)$, we can modify the strategy $(b,b) = (\infty,\infty)$ in the same way as above (for a sufficiently large b') to find a strategy that outperforms $\boldsymbol{b}$.

A.7 Proof of Corollary 7

Let $\hat{V}_{\infty}^{\bar{p},\bar{q}}(b_{im}, b_{in})$ denote the expression $\hat{V}_{\infty}(b_{im}, b_{in})$ when the parameter choices $p = \bar{p}$ and $q = \bar{q}$ are made. Similarly, let $\hat{V}_{\infty}^{\bar{p},\bar{q},\bar{L}_{im,\infty},\bar{L}_{in,\infty}}(b_{im}, b_{in})$ denote the expression $\hat{V}_{\infty}(b_{im}, b_{in})$ when the parameter choices $p = \bar{p}$, $q = \bar{q}$, $L_{im,\infty} = \bar{L}_{im,\infty}$ and $L_{in,\infty} = \bar{L}_{in,\infty}$ are made. When parameters $L_{im,\infty}$ and $L_{in,\infty}$ are omitted from the superscript, the meaning is that they are assumed to be the original fixed ones.

For each of part (a) and part (b), we will show a stronger statement than the theorem statement. Specifically, we will show that for any pair of decreasing sequences $\{p_n\}_{n\in\mathbb{N}}$ and $\{q_m\}_{m\in\mathbb{N}}$ converging to zero, there exists N such that for any $n \ge N$, we can find M_n such that the quitting point of innovation-learning tasks b_{in} of any strategy $\boldsymbol{b} = (\infty, b_{in})$ maximizing $V_{\infty}^{p_n,q_m}$ is greater than γ for all $m \ge M_n$.

We note that the sequence of continuous functions $\{\hat{V}_{\infty}^{p_n,0}\}_{n\in\mathbb{N}}$ pointwise converge to the continuous function $\hat{V}_{\infty}^{0,0}$. By part (a) of Lemma 8, this sequence of continuous functions in fact monotonically converges (increasing with respect to n) to $\hat{V}_{\infty}^{0,0}$. An application of Dini's theorem thus shows that the convergence of $\{\hat{V}_{\infty}^{p_n,0}\}_{n\in\mathbb{N}}$ to $\hat{V}_{\infty}^{0,0}$ on the compact space $\bar{\mathcal{Q}} \cup \{(0,0)\}$ is uniform.

The proof of Proposition 5 implies that the maximum of $\hat{V}_{\infty}^{0,0}$ on $\bar{\mathcal{Q}} \cup \{(0,0)\}$ is attained at $(b_{im}, b_{in}) = (b_{\text{arbitrary}}, \infty)$; note that the subscript "arbitrary" means the choice of that parameter has no effect. Indeed, check that

$$\hat{V}_{\infty}^{0,0}(b_{\text{arbitrary}}, b) = \hat{V}_{\infty}^{p_{\text{arbitrary}},1,L_{in,\infty},L_{in,\infty}}(b, b'_{\text{arbitrary}}), \tag{245}$$

since when $p = 0$, we have the equality of distributions $\mu_{in} = \mu_{im}$; and we have set the learning function of imitation-learning tasks to be $L_{in,\infty}$ as well. The proof of Proposition 5 shows that we have

$$\hat{V}_{\infty}^{p_{\text{arbitrary}},1,L_{in,\infty},L_{in,\infty}}(b, b) < \hat{V}_{\infty}^{p_{\text{arbitrary}},1,L_{in,\infty},L_{in,\infty}}(\infty, b), \tag{246}$$

where we have set $b'_{\text{arbitrary}} = b$. This shows that $(\infty, b'_{\text{arbitrary}})$ maximizes the function $\hat{V}_{\infty}^{p_{\text{arbitrary}},1,L_{in,\infty},L_{in,\infty}}$; equivalently, $(b_{\text{arbitrary}}, \infty)$ maximizes the function $\hat{V}_{\infty}^{0,0}$.

We will now avoid the use of the term $b_{\text{arbitrary}}$, and instead define

$$\tilde{V}(b) = \hat{V}_{\infty}^{0,0}(b_{\text{arbitrary}}, b). \tag{247}$$

Let $\gamma \geq 0$. Consider the positive number

$$\varepsilon = \tilde{V}(\infty) - \max_{b \leq \gamma} \tilde{V}(b). \tag{248}$$

By uniform convergence, there exists N such that for all $n \geq N$, we simultaneously have

$$|\hat{V}_{\infty}^{p_n,0}(b_1, \infty) - \hat{V}_{\infty}^{0,0}(b_1, \infty)| < \frac{\varepsilon}{2} \tag{249}$$

and

$$|\hat{V}_{\infty}^{p_n,0}(b_2, b) - \hat{V}_{\infty}^{0,0}(b_2, b)| < \frac{\varepsilon}{2} \tag{250}$$

for all $b \leq \gamma$. Since

$$\hat{V}_{\infty}^{0,0}(b_1, b) = \tilde{V}(b) < \tilde{V}(\infty) < \hat{V}_{\infty}^{0,0}(b_2, \infty) \tag{251}$$

for every $b \leq \gamma$ with a difference of at least ε, it follows from the triangle inequality that for any

$n \geq N$, we have

$$\hat{V}_{\infty}^{p_n,0}(b_2, b) < \hat{V}_{\infty}^{p_n,0}(b_1, \infty) \tag{252}$$

for all $b \leq \gamma$. Note that the choice of b_1 and b_2 has no effect on the values $\hat{V}_{\infty}^{p_n,0}(b_2, b)$ and $\hat{V}_{\infty}^{p_n,0}(b_1, \infty)$.

Fix $n \geq N$. By part (b) of Lemma 8, the continuous functions $\{\hat{V}_{\infty}^{p_n,q_m}\}_{m\in\mathbb{N}}$ monotonically converge (decreasing with respect to m) to $\hat{V}_{\infty}^{p_n,0}$, which is also continuous. It thus follows from Dini's theorem that the convergence of $\{\hat{V}_{\infty}^{p_n,q_m}\}_{m\in\mathbb{N}}$ to $\hat{V}_{\infty}^{p_n,0}$ on the compact space $\bar{\mathcal{Q}} \cup \{(0,0)\}$ is uniform. Let

$$\varepsilon = \hat{V}_{\infty}^{p_n,0}(b_1, \infty) - \sup_{b\leq\gamma} \hat{V}_{\infty}^{p_n,0}(b_2, b), \tag{253}$$

which is positive by our choice of n. By uniform convergence, there exists M_n such that for all $m \geq M_n$, we simultaneously have the inequality

$$\left| \hat{V}_{\infty}^{p_n,q_m}(\infty, \infty) - \hat{V}_{\infty}^{p_n,0}(\infty, \infty) \right| < \frac{\varepsilon}{2}, \tag{254}$$

where we have set $b_1 = \infty$; and the inequality

$$\left| \hat{V}_{\infty}^{p_n,q_m}(b_2, b) - \hat{V}_{\infty}^{p_n,0}(b_2, b) \right| < \frac{\varepsilon}{2} \tag{255}$$

for any $(b_2, b) \in \bar{\mathcal{Q}}$. Since

$$\hat{V}_{\infty}^{p_n,0}(b_2, b) < \hat{V}_{\infty}^{p_n,0}(\infty, \infty) \tag{256}$$

for all $b \leq \gamma$ with a difference of at least ε, it follows from the triangle inequality that for any $m \geq M_n$, we have

$$\hat{V}_{\infty}^{p_n,q_m}(b_2, b) < \hat{V}_{\infty}^{p_n,q_m}(\infty, \infty) \tag{257}$$

for all $b \leq \gamma$. Setting $b_2 = \infty$ yields part (b), while setting $b_2 = b$ yields part (a).

In this proof, we have applied the proof of Proposition 5 to show a necessary lemma that can be described as the following. In the continuous learning model with parameters $q = 1$ and $L_{im,\infty}(t) = L(t)$, the unique strategy to maximize $V_{\infty}(\boldsymbol{b})$ is $\boldsymbol{b} = \infty$, where the single entry denotes that there is no ambiguity in learning types. Strictly speaking, the proof of Proposition 5 applied to this continuous learning game only shows that $\boldsymbol{b} = \infty$ outperforms $\boldsymbol{b}' = b \in (0, \infty)$, and not necessarily $b \to 0$. This leaves the possibility that $V_{\infty}(b)$ is also maximized at $b \to 0$, with the same function value as $b = \infty$. This implies that V_{∞} is decreasing near $b = 0$. However, a quick application of the proof of Proposition 5 shows that this possibility does not arise. Specifically, this proof shows that for small $b > 0$, the strategy $\boldsymbol{b}' = b$ is strictly outperformed by

$$\boldsymbol{b}'' = \left(L\left(2L^{-1}(b)\right), b, b, \ldots\right), \tag{258}$$

which—as a subsequent application of this proof shows—is strictly outperformed by

$$\boldsymbol{b}''' = \left(L\left(2L^{-1}(b)\right), L\left(2L^{-1}(b)\right), b, b, \ldots\right). \tag{259}$$

Continuing iteratively, we obtain that $\boldsymbol{b}'$ is strictly outperformed by the strategy

$$\hat{\boldsymbol{b}} = L\left(2L^{-1}(b)\right). \tag{260}$$

Taking b to be small, we see that $\tilde{V}(0) = \lim_{b\to 0} V(b)$ cannot be decreasing near $b = 0$, and thus it is impossible that the maximum is attained at the two endpoints $b = 0$ and $b = \infty$.

A.8 Proof of Lemma 8

Recall that

$$\hat{V}_\infty(b_{im}, b_{in}) = q\hat{V}_{im}(b_{im}, b_{in}) + (1-q)\hat{V}_{in}(b_{im}, b_{in}) \tag{261}$$

for

$$\hat{V}_{im}(b_{im}, b_{in}) = \frac{\mathfrak{d}\mathfrak{e} - \mathfrak{b}\mathfrak{f}}{\mathfrak{g}} \tag{262}$$

and

$$\hat{V}_{in}(b_{im}, b_{in}) = \frac{\mathfrak{a}\mathfrak{f} - \mathfrak{c}\mathfrak{e}}{\mathfrak{g}}. \tag{263}$$

First, we show that

$$\frac{\partial}{\partial p}\hat{V}_\infty(b_{im}, b_{in}) \leq 0, \tag{264}$$

with equality if and only if $q = 1$. Note that the only one of $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}, \mathfrak{e}, \mathfrak{f}$, and $\mathfrak{g}$ that is not constant with respect to p is $\boldsymbol{f}$. We thus have

$$\frac{\partial}{\partial p}\hat{V}_\infty(b_{im}, b_{in}) = \frac{\partial}{\partial p}\hat{V}_\infty(b_{im}, b_{in}) = \frac{-q\mathfrak{b} + (1-q)\mathfrak{a}}{\mathfrak{g}}\left(\frac{\partial}{\partial p}\mathfrak{f}\right). \tag{265}$$

Check that

$$-q\mathfrak{b} + (1-q)\mathfrak{a} = q(1-q)\delta^{L^{-1}_{in,\infty}(b_{in})}\eta^{b_{im}} + 1 - q - q(1-q)\delta^{L^{-1}_{im,\infty}(b_{im})}\eta^{b_{im}}$$

$$
\begin{aligned}
&\geq 1 - q - q(1-q)\delta^{L_{im,\infty}^{-1}(b_{im})}\eta^{b_{im}} \\
&\geq 1 - q - q(1-q) = (1-q)^2, \qquad (266)
\end{aligned}
$$

which is nonnegative, and positive if and only if $q < 1$. Check also that

$$
\boldsymbol{g} = 1 - q - (1-q)\delta^{L_{in,\infty}^{-1}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right) + q - q\delta^{L_{im,\infty}^{-1}(b_{im})}\eta^{b_{im}} \geq 0. \qquad (267)
$$

with equality if and only if $(b_{im}, b_{in}) = (0,0)$, since

$$
\delta^{L_{in,\infty}^{-1}(b_{in})}\left(p + (1-p)\eta^{b_{in}}\right), \delta^{L_{im,\infty}^{-1}(b_{im})}\eta^{b_{im}} \leq 1. \qquad (268)
$$

Then, check that

$$
\begin{aligned}
\frac{\partial}{\partial p}\mathfrak{f} = \frac{\partial}{\partial p}\Bigg(p \cdot 0 + (1-p)\Bigg(&\int_0^{b_{in}}\left(\int_0^{L_{in,\infty}^{-1}(a)}\delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^{\infty}\delta^t dt\right)d\mu_\eta(a) \\
&+ \int_{a>b_{in}}\left(\int_0^{L_{in,\infty}^{-1}(b_{in})}\delta^t f(a, L_{in,\infty}(t))dt\right)d\mu_\eta(a)\Bigg)\Bigg) \\
= -\Bigg(&\int_0^{b_{in}}\left(\int_0^{L_{in,\infty}^{-1}(a)}\delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^{\infty}\delta^t dt\right)d\mu_\eta(a) \\
&+ \int_{a>b_{in}}\left(\int_0^{L_{in,\infty}^{-1}(b_{in})}\delta^t f(a, L_{in,\infty}(t))dt\right)d\mu_\eta(a)\Bigg) \leq 0, \qquad (269)
\end{aligned}
$$

with equality if and only if $b_{in} = 0$, which is equivalent to $(b_{im}, b_{in}) = (0,0)$ in our domain $\bar{\mathcal{Q}} \cup \{(0,0)\}$. Finally, it follows from the calculation via L'Hôspital's rule in the proof of Proposition 4 that

$$
\frac{\frac{\partial}{\partial p}\mathfrak{f}}{\mathfrak{g}} \to \frac{\log\frac{1}{\eta}}{\left(\log\frac{1}{\delta}\right)\frac{d}{db}\mathfrak{g}(b,b)|_{b=0}} > 0 \qquad (270)
$$

as $b \to 0$ for $(b_{im}, b_{in}) = (b,b)$. The condition $(b_{im}, b_{in}) = (0,0)$ does not make (265) zero.

We thus have shown (264), where equality holds if and only if $q = 1$.

Next, suppose that Assumption 1 holds and the imitation-learning knowledge function $L_{im,\infty}(t)$ is convex. We need to show that

$$
\frac{\partial}{\partial q}\hat{V}_\infty(\infty, b_{in}) > 0. \qquad (271)
$$

Note that $\mathfrak{e}$ and $\mathfrak{f}$ are constant in q, while $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \mathfrak{d}$, and $\mathfrak{g}$ are not. Check that at $(b_{im}, b_{in}) =$

(∞, b_{in}), we have

$$\mathfrak{a} = 1, \tag{272}$$

$$\mathfrak{b} = 0, \tag{273}$$

$$\mathfrak{c} = -q\delta^{L^{-1}_{in,\infty}(b_{in})}(p + (1-p)\eta^{b_{in}}), \tag{274}$$

and

$$\mathfrak{d} = 1 - (1-q)\delta^{L^{-1}_{in,\infty}(b_{in})}(p + (1-p)\eta^{b_{in}}). \tag{275}$$

Let

$$\mathfrak{h} = \delta^{L^{-1}_{in,\infty}(b_{in})}(p + (1-p)\eta^{b_{in}}). \tag{276}$$

We next apply the quotient rule to obtain

$$\begin{aligned}
&\frac{\partial}{\partial q}\hat{V}_\infty(\infty, b_{in}) \\
&\quad = \frac{1}{\mathfrak{g}^2}\left(\mathfrak{g}\frac{\partial}{\partial q}\left(q\mathfrak{d}\mathfrak{e} + (1-q)\mathfrak{f} - (1-q)\mathfrak{c}\mathfrak{e}\right) - \left(q\mathfrak{d}\mathfrak{e} + (1-q)\mathfrak{f} - (1-q)\mathfrak{c}\mathfrak{e}\right)\frac{\partial \mathfrak{g}}{\partial q}\right) \\
&\quad = \frac{(1-(1-q)\mathfrak{h})(\mathfrak{e}-\mathfrak{f}) - \mathfrak{h}(\mathfrak{f} + q(\mathfrak{e}-\mathfrak{f}))}{(1-(1-q)\mathfrak{h})^2} \\
&\quad = \frac{\mathfrak{e} - \mathfrak{f} - \mathfrak{e}\mathfrak{h}}{(1-(1-q)\mathfrak{h})^2}.
\end{aligned}$$

If $b_{im} = \infty$ so that $\mathfrak{h} = 0$, then we have

$$\frac{\partial}{\partial q}\hat{V}_\infty(\infty, b_{in}) = \mathfrak{e} - \mathfrak{f}, \tag{277}$$

which is positive; indeed, check that

$$\mathfrak{e} = \int_0^\infty \left(\int_0^{L^{-1}_{im,\infty}(a)} \delta^t f(a, L_{im,\infty}(t))dt + \int_{L^{-1}_{im,\infty}(a)}^\infty \delta^t dt\right) d\mu_\eta(a), \tag{278}$$

while

$$\mathfrak{f} = (1-p)\int_0^\infty \left(\int_0^{L^{-1}_{in,\infty}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L^{-1}_{in,\infty}(a)}^\infty \delta^t dt\right) d\mu_\eta(a). \tag{279}$$

We see that

$$
\begin{aligned}
&(1-p)\int_0^\infty \left(\int_0^{L_{in,\infty}^{-1}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^\infty \delta^t dt\right) d\mu_\eta(a) \\
&< \int_0^\infty \left(\int_0^{L_{in,\infty}^{-1}(a)} \delta^t f(a, L_{in,\infty}(t))dt + \int_{L_{in,\infty}^{-1}(a)}^\infty \delta^t dt\right) d\mu_\eta(a) \\
&\leq \int_0^\infty \left(\int_0^{L_{im,\infty}^{-1}(a)} \delta^t f(a, L_{im,\infty}(t))dt + \int_{L_{im,\infty}^{-1}(a)}^\infty \delta^t dt\right) d\mu_\eta(a), \qquad (280)
\end{aligned}
$$

as needed, since $L_{in,\infty}(t) \leq L_{im,\infty}(t)$.

Now, suppose that $b_{im} < \infty$, so that $\mathfrak{h} > 0$. Then, we can write

$$\frac{\partial}{\partial q}\hat{V}_\infty(\infty, b_{in}) = \frac{\mathfrak{e} - \mathfrak{f} - \mathfrak{e}\mathfrak{h}}{(1-(1-q)\mathfrak{h})^2} = \frac{\hat{V}_{im}(\infty, b_{in}) - \hat{V}_{in}(\infty, b_{in})}{1-(1-q)\mathfrak{h}}, \qquad (281)$$

since

$$\hat{V}_{im}(\infty, b_{in}) - \hat{V}_{in}(\infty, b_{in}) = \frac{\mathfrak{d}\mathfrak{e} - \mathfrak{b}\mathfrak{f} - (\mathfrak{a}\mathfrak{f} - \mathfrak{c}\mathfrak{e})}{\mathfrak{g}} = \frac{\mathfrak{e} - \mathfrak{f} - \mathfrak{e}\mathfrak{h}}{1-(1-q)\mathfrak{h}}. \qquad (282)$$

Thus, we need to show that

$$\hat{V}_{im}(\infty, b_{in}) > \hat{V}_{in}(\infty, b_{in}). \qquad (283)$$

This inequality is proven by showing that

$$\hat{V}_{im}(\infty, b_{in}) > \hat{V}_{im}(\boldsymbol{b}) > \hat{V}_{in}(\infty, b_{in}), \qquad (284)$$

for

$$\boldsymbol{b} = \left((L_{im,\infty}(L_{in,\infty}^{-1}(b_{in})), b_{in}), (\infty, b_{in}), (\infty, b_{in}), \ldots\right). \qquad (285)$$

The comprising inequality

$$\hat{V}_{im}(\boldsymbol{b}) < \hat{V}_{im}(\infty, b_{in}) \qquad (286)$$

follows from the optimality of never quitting tasks of type $j = im$, demonstrated in the proof of Proposition 5. Indeed, conditional on the current task being of type $j = im$, the strategy (∞, b_{in}) is equivalent to never quitting this curren task.

Only the comprising inequality

$$\hat{V}_{in}(\infty, b_{in}) < \hat{V}_{im}(\boldsymbol{b}). \qquad (287)$$

remains to be shown. The sample space of task sequences for the left-hand-side value function $\hat{V}_{im}(\boldsymbol{b}, b_{in})$ is

$$(0, \infty) \times \mathcal{U}^{\infty} \tag{288}$$

with the probability measure

$$\mu_{im} \otimes \mu^{\infty} = \mu_{\eta} \otimes \mu^{\infty}, \tag{289}$$

and the sample space of task sequences of the right-hand-side value function $\hat{V}_{in}(b_{im}, b_{in})$ is

$$((0, \infty) \cup \{\infty\}) \times \mathcal{U}^{\infty} \tag{290}$$

with the probability measure

$$\mu_{in} \otimes \mu^{\infty}, \tag{291}$$

where we recall that μ_{in} places probability p on $a = \infty$ and distributes the remaining probability $1 - p$ as the exponential distribution μ_{η}. It suffices to show that

$$\begin{aligned} &\int_{a_1 \in (0,\infty)} \left(\int_{(j_2,a_2),\ldots)\in\mathcal{U}^{\infty}} V_{in}((\infty, b_{in}), ((in, a_1), (j_2, a_2), \ldots) d\mu^{\infty} \right) d\mu_{\eta} \\ &\qquad \leq \int_{a_1 \in (0,\infty)} \left(\int_{((j_2,a_2),\ldots)\in\mathcal{U}^{\infty}} V_{im}(\boldsymbol{b}, ((im, a_1), (j_2, a_2), \ldots)) d\mu^{\infty} \right) d\mu_{\eta} \end{aligned} \tag{292}$$

and

$$\begin{aligned} &\int_{a_1 \in \{\infty\}} \left(\int_{(j_2,a_2),\ldots)\in\mathcal{U}^{\infty}} V_{in}((\infty, b_{in}), ((in, a_1), (j_2, a_2), \ldots) d\mu^{\infty} \right) d\chi \\ &\qquad < \int_{a_1 \in (0,\infty)} \left(\int_{((j_2,a_2),\ldots)\in\mathcal{U}^{\infty}} V_{im}(\boldsymbol{b}, ((im, a_1), (j_2, a_2), \ldots)) d\mu^{\infty} \right) d\mu_{\eta} \end{aligned} \tag{293}$$

for χ the one-point distribution on $\{\infty\}$. Indeed, adding the product of the inequality (292) with $(1 - p)$ with the product of the inequality (293) with p yields the desired inequality (287).

The second inequality (293) holds immediately because it simplifies to

$$0 < \int_{a_1 \in (0,\infty)} \left(\int_{((j_2,a_2),\ldots)\in\mathcal{U}^{\infty}} V_{im}(\boldsymbol{b}, ((im, a_1), (j_2, a_2), \ldots)) d\mu^{\infty} \right) d\mu_{\eta}. \tag{294}$$

The first inequality (292) holds because for every sample

$$(a_1, (j_2, a_2), \ldots), \tag{295}$$

the payoff of the left-hand-side value function

$$V_{in}((\infty, b_{in}), ((in, a_1), (j_2, a_2), \ldots)) \tag{296}$$

is at most the payoff of the right-hand-side value function

$$V_{im}(\boldsymbol{b}, ((im, a_1), (j_2, a_2), \ldots)). \tag{297}$$

This is demonstrated by partitioning $[0, \infty)$ into various subintervals and looking at the respective sub-payoff values corresponding to each subinterval.

In the subinterval

$$[0, L^{-1}_{in,\infty}(b_{in})), \tag{298}$$

the sub-payoff value of the left-hand-side value function is at most that of the right-hand-side value function, because the learning of the first task is faster for the latter than the former: $L_{in,\infty}(t) \leq L_{im,\infty}(t)$.

In the subinterval

$$[L^{-1}_{in,\infty}(b_{in}), \infty), \tag{299}$$

the sub-payoff value of the left-hand-side value function is equal to that of the right-hand-side value function conditional on the first task being quit at time $t = L^{-1}_{in,\infty}(b_{in})$ for both, i.e., conditional on learning not yet having completed. And conditional on the opposite—that learning of the first task completes for at least one of the value functions by time $t = L^{-1}_{in,\infty}(b_{in})$—we have the following. If this occurs for the left-hand-side value function, then it also occurs for the right-hand side value function, since $L_{in,\infty}(t) \leq L_{im,\infty}(t)$. Thus, we have the desired inequality for the sub-payoff values corresponding to the subinterval (299). If this occurs for the right-hand-side value function, then its sub-payoff value corresponding to the subinterval (299) is maximal, so the inequality holds anyway. Thus, we have obtained (293), and thereby the desired inequality (287).

This concludes our proof of (271).