Strategic variability in humans, pigeons, and rats.

Janne Reynders, Tom Verguts, & Senne Braem

Department of Experimental Psychology, Ghent University, Belgium

Corresponding author: Janne Reynders; janne.reynders@ugent.be

Acknowledgements

This work was supported by an ERC Starting grant awarded to S.B. (European Union's Horizon 2020 research and innovation program, Grant agreement 852570). We would like to thank Greg Jensen and Allen Neuringer for providing us with their data and useful discussions prior to our study.

Abstract

Strategically variable behavior can be advantageous in various fields such as sports (unpredictability), art (creativity), science (innovation), and problem-solving (thinking outside the box). Although previous studies identified experimental conditions under which humans and non-human animals show increased variable decision-making, we have only a limited understanding of its underlying cognitive mechanisms. Using a reinforcement learning model, we simulate the use of three different theorized strategies in an adversarial reward learning environment that requires very high variability. Model simulations with a policy-gradient meta-learning algorithm show that agents could respond more optimally in such environments by (1) relying on a stochastic generator, (2) increasing one's learning rate to allow for faster interactions between reinforcement learning and extinction, or (3) strategically upvalue unchosen actions using a frequency-based memory. After demonstrating the theoretical benefit of each of these strategies, we fitted our model on existing datasets of human-, pigeons- and rat behavior in adversarial environments. We show that, while all three species can engage in highly variable behavior, only humans strategically upvalue unchosen actions as a strategy to achieve variability.

Introduction

The ability to behave variable, unpredictable, or seemingly random can often be a beneficial strategy. For example, a football player will benefit from being unpredictable during shootouts, to avoid that the goalkeeper can predict the ball's trajectory (S. Gershman, 2021). Prey animals show random-like behavior when chased by a predator to confuse or disorient the predator (Humphries & Driver, 1967). More generally, variable behavior can yield alternative ways to reach a goal, lead to new ideas for solving a problem, or be innovative in science, creative in art, or unpredictable in games (Campbell, 1960; Kilicay-Ergin & Jablokow, 2012; Parsonson & Baer, 1978; Walker & Wooders, 2001). It improves the ability to adjust to new, uncertain or complex environments (Hills et al., 2015). Without regular sources of variation, action selection would be confined to information extracted from prior learning episodes, making progress difficult (Nergaard & Holth, 2020; Uddin, 2021). On a larger time scale, variability drives individual development, everyday learning, societal advancement and natural evolution (Dall & Griffith, 2014; Donahoe & Palmer, 1994; Siegler, 1998). Despite this broad agreement on the benefits of variability for learning and action and the numerous examples where humans and non-human animals act strategically variable, it remains unclear *how* humans and other animals make variable decisions. Specifically, we lack insight into the cognitive processes that underlie the generation of *strategically* variable behavior.

Decision-making in biological and artificial agents alike fundamentally involves the exploitation-exploration trade-off; the choice to exploit familiar options for a known reward or to instead explore new options for an unknown, potentially better, reward (Mehlhorn et al., 2015). The latter is sometimes implemented using novelty bonuses (Kakade & Dayan, 2002). Exploration is even more important when variability *itself* is rewarded. To test such strategic variability, some

studies in the reinforcement learning literature have used reinforcement schedules that are aimed to increase action variability. These experiments typically involve a multi-armed bandit problem where on each trial subjects have to choose among several alternatives, and response variability is encouraged by only rewarding the previously least frequent and/or recent (sequence of) chosen actions (Page & Neuringer, 1985). Hence, rewards are maximized when subjects respond as variable as possible. From here on, we will refer to such an environment as adversarial, as the environment attempts to oppose the subject and drive them to a sequence of variable actions as a result. Such adversarial environments successfully increase choice variability, including in humans (Jensen et al., 2006; Neuringer, 1986), pigeons (Jensen et al., 2006; Machado, 1989, 1992, 1993; Page & Neuringer, 1985) and rats (Bryantt & Church, 1974; Neuringer, 1991).

These experiments show that humans and non-human animals can learn to behave variably when the environment requires them to do so. However, the cognitive mechanisms underlying this strategic variability remain unknown. Several ideas have been proposed from different cognitive theories. The current study will formalize three different proposed cognitive mechanisms within one computational modelling framework, investigate their respective computational efficiency, and empirical fit to different data. This will be achieved by associating each of the three cognitive mechanisms to one critical parameter in the model and its dynamics across various environments. We will refer to the potential cognitive mechanisms as: a stochastic-generator, dynamic reinforcement/extinction learning, and a frequency-based memory. Before turning to our model, we briefly introduce each of these putative mechanisms in turn.

The first proposal is that the brain has access to a *stochastic*[1] *generator* (Jensen et al., 2006; Page & Neuringer, 1985): a mechanism that allows humans and non-human animals to select random actions, i.e., without regard to any learned information. Learning from past experience and previous choices would still happen, but this information can be blocked when variability is required. Similarly, Maye et al., (2007) propose that 'a general neural mechanism' underlies spontaneous behavior. Jensen et al., (2006) conducted experiments on humans and pigeons, where on each trial, subjects had to make one choice among several alternatives. An adversarial environment was used to induce variability in the subjects' actions. They exposed their subjects to three different experimental conditions, varying the number of alternative choice options in each condition (that is, two, four or eight alternatives). They argued that, if subjects use a stochastic generator, performance should improve with *more* action options, because a random generator is less likely to repeat a sequence with more options. On the other hand, if performance were to decrease with increasing alternatives, they hypothesized that the subjects would rely on more a systematic approach to being variable, as it's easier to use a systematic process when there are less options to choose from. They observed an increase in performance when increasing the choice alternatives and concluded that this is consistent with a stochastic-generator hypothesis.

The idea of a stochastic generator is also implicit in computational models of decision-making and reinforcement learning (Sutton & Barto, 2020). Decision making requires dealing with the exploration-exploitation trade-off, and this can be done by introducing parameters that quantify

[1] With variability, we refer to a broader concept that relates to a property of the desired or observed behavior. Stochasticity is one way of achieving variability. However, behavior can be highly variable without a random underlying process.

the amount of randomness that goes into the decision process algorithm. One such parameter is the epsilon parameter in the epsilon-greedy algorithm (Abed-alguni, 2018; Sutton & Barto, 2020; Wilson & Collins, 2019). Epsilon represents the probability to select an action randomly, and hence, explore: the higher epsilon, the more actions are chosen randomly. Hence, fitting such a model to behavior implicitly assumes that humans and non-human animals have access to a stochastic generator.

Dynamic reinforcement/extinction learning provides a second potential mechanism that could lead to behavioral variability (Machado & Tonneau, 2012; Nergaard & Holth, 2020). Here, the agent would quickly learn from previous feedback and prediction errors: infrequent actions are more likely to lead to reinforcement, making them more frequent, while frequently chosen options remain unrewarded and lead to extinction, making them infrequent (Machado & Tonneau, 2012). Applying this reinforcement loop at a very high pace (e.g., immediate reinforcement of individual actions when rewarded and immediate extinction when unrewarded), could ultimately also result in the generation of highly variable action sequences. In models of decision-making, this could be achieved with a very high learning rate in Rescorla-Wagner style update rules. In this case, agents would learn to behave variably based on the outcome of an action or choice.

A third potential mechanism relies on a *frequency-based memory*. Such a mechanism can lead to variability by adapting action values as a function of how (in)frequently they were selected or, in other words, an agent would store the actions that were previously taken and then decide to not do the exact same thing again. In practice, this requires remembering and monitoring previous actions and track their individual choice frequencies. Storing this information can allow choosing infrequent actions and thus avoid choosing the same actions again. Like the stochastic generator, this strategy can result in highly variable behavior, but it originates from a different cognitive

process, i.e., by relying on past choices. It requires memory to store choice frequencies of past actions and a choice is made based on how often the option was chosen in the past, irrespective of their outcome as would be the case of dynamic reinforcement/extinction learning. This idea is consistent with the common interpretation of random behavior in the random number generation (RNG) task (Ross, 1955), where people are asked to produce a random sequence of numbers. In the RNG literature, it is commonly assumed that the underlying process for generating variable behavior is related to a combination of executive functions, inhibition and working memory strategies that rely on the tracking and remembering of the frequency of omitted actions (Baddeley et al., 1998; Capone et al., 2014; Joppich et al., 2004; Oomens et al., 2015; Towse, 1998; Wagenaar, 1972).

Here, we first suggest a computational modeling framework that integrates these three potential cognitive sources of variability. This allows us to distinguish the respective contributions of each of these three proposed mechanisms, and thus eventually elucidate which (combination of) processes are used to produce strategically variable behavior in different species and contexts. For this purpose, we adapt one of the most influential models of reinforcement learning and decision-making – the Rescorla-Wagner model (Rescorla & Wagner, 1972), and link each mechanism to one of three parameters in our model.

The model

The modeling framework is based on a standard Rescorla-Wagner (RW) model for reinforcement learning and decision-making (Rescorla & Wagner, 1972). Each response option or action is associated to a *Q*-value that is updated on a trial-by-trial basis according to a learning rule (defined shortly). On each trial, decision-making relies on an epsilon-greedy decision-making rule,

where epsilon quantifies the level (probability) of stochasticity in the decisions. Otherwise, the action with the currently highest (learned) *Q*-value is chosen:

$$\text{Decision at trial } t\text{: } \begin{cases} action\ with\ argmax\{Q_{t-1}(a)\} & with\ P = 1-\varepsilon \\ random\ action & with\ P = \varepsilon \end{cases} \quad (1)$$

where *a* stands for action, $Q_t(a)$ is the *Q*-value at trial *t* for action *a* and *P* stands for probability. The option with the currently highest *Q*-value is selected with a probability of 1 - ε. In the other case, an option is randomly selected. In other words, ε represents the probability to select an option randomly, without taking past information (based on learned *Q*-values) into account. Hence, low and high values of ε lead to stable and variable decisions, respectively. The stochastic generator mechanism predicts a regulation of ε across environments that require different levels of behavioral- or decision variability.

We opted for an epsilon-greedy decision-making policy instead of another commonly used function to introduce randomness, i.e., the temperature parameter in a Softmax function, because we wanted a parameter that directly describes the proportion of completely random choices. Specifically, our aim was to have a simple policy with only two possibilities: either select the action with the highest learned value (the greedy choice) or make a completely random choice. Therefore, our approach demands a parameter that strictly reflects the probability of making a completely random decision, without any dependence on past learned values. In contrast, a Softmax decision policy still includes the relative values of the choice options, meaning it relies on past information even when making probabilistic choices.

After the choice is made and feedback is given, the *Q*-value of the currently chosen option is updated according to a RW learning rule. Specifically:

$$Q_t(a,c) = Q_{t-1}(a,c) + \alpha\big(r_{t-1} - Q_{t-1}(a,c)\big) \quad (2)$$

Here *a* stands for action, *c* stands for chosen, α is the learning rate, *r* is the reward and *t* is the trial. The *Q*-value of the chosen action is updated with the prediction error weighted by a learning rate α. The prediction error, $r_{t-1} - Q_{t-1}(a, c)$, is the difference between the value of the reward and the previous *Q*-value of the currently chosen option. The learning rate α quantifies the speed with which updates are made after feedback. The dynamic reinforcement/extinction learning mechanism predicts that an increase in learning rate is the prerequisite to generate variable behavior.

Finally, the standard RW learning rule for chosen options was extended with an update rule for all remaining, unchosen options. We added a parameter, λ, that allows to keep track of the frequency with which an action wasn't chosen through its *Q*-value.

$$Q_t(a, u) = Q_{t-1}(a, u) + \lambda \quad (3)$$

Here, *u* stands for unchosen and λ is a value-bias towards recently unchosen options, further referred to as the unchosen value-bias. This parameter updates the frequency-based memory. The less frequently a particular option is chosen, the more λ is added to its associated *Q*-value. This can upregulate (if $\lambda > 0$) or downregulate (if $\lambda < 0$) the value of unchosen actions, leading to higher (resp. lower) *Q*-values for unchosen options, making them more (resp. less) likely to be chosen. This can result in more variable (resp. more stable) choices relative to standard RW-based learning. The frequency-based memory mechanism predicts an upregulation of λ when more strategically variable behavior is generated.

Taken together, our model has three critical parameters, ε, α, and λ, each of which closely align with one of the three cognitive mechanisms introduced above. Each of these mechanisms predicts the upregulation of a different parameter. That is, the stochastic generator mechanism

predicts the upregulation of ε, the dynamic reinforcement/extinction learning mechanism predicts the upregulation of α, and the frequency-dependent memory mechanism predicts the upregulation of λ, when strategically variable behavior is required.

This paper presents two studies. Study 1 is a computational investigation that tested the above predictions by simulating the model across different environments that require different levels of variable responding. This allowed us to evaluate the internal validity of our model. Specifically, the model was simulated in a stable environment (where the different response options have fixed reward probabilities), a volatile environment (where the reward probabilities are shuffled among the different response options every few trials), and, most importantly, an adversarial environment (following a variability contingency where only the 60% least frequent and least recent response duplets are rewarded). These three environments have an increasing demand to make variable decisions in order to get rewards. A policy-gradient optimization of ε, α, and λ (in separate simulations) was used to find the optimal parameter values in each environment and to investigate whether adjusting any of these three parameters across the environments is beneficial to adapt to the different variability demands, and in particular the adversarial one. For all three parameters, we predicted that higher values would allow for more variable behavior.

Next, Study 2 is an empirical investigation that evaluated the applicability and external validity of this model by fitting it to behavioral data from humans, pigeons and rats exposed to an adversarial context. We predicted that the preferred mechanism(s) employed by humans pigeons and rats would be reflected by higher values of these parameters when fitting the model, and compared parameter values across species to see whether cross-species differences could be observed in generating strategically variable behavior.

Study 1: Model simulations across stable, volatile, and adversarial environments

Methods

In this Simulation study, the model performed a multi-armed bandit task in a stable, a volatile, and an adversarial environment to investigate how it would adapt its parameters in each. All environments contained eight choice options. In the stable environment, three options had a probability of 70% to yield a reward and the remaining five had a 30% reward probability. These reward probabilities remained fixed across trials. In the volatile environment, three options had a 90% reward probability and the remaining five a 10% reward probability. Importantly, these reward probabilities shuffled on average every 15 trials (taken from a rounded normal distribution $\mathcal{N}(15,3)$). Finally, the adversarial environment also included eight choice options, but reinforcement was based on choice frequencies (Machado, 1992; Page & Neuringer, 1985). Frequencies of response pairs (the current and previous response) were tracked. This resulted in 64 ($= 8^2$) different response pairs, that were all initialized at a counter uniformly randomly sampled between 0.9 and 1.1. After each choice (except for the very first choice), the counter of the corresponding response pair consisting of the current choice and the previous choice was increased by one. All other counters were subtracted by 1/63. Finally, all counters were multiplied by an amnesia coefficient of $\varphi = 0.984$ to account for recency (see also Denney & Neuringer, 1998). This ensured that the less recent a response pair was emitted, the more its associated counter was decreased, making less recent pairs more likely to be reinforced. Reinforcement was given when the current completed response pair was associated to a counter that was among the 60% lowest counters. For example, if the response options are numbered from 1 to 8, consider 15 trials where

the sequence of choices thus far was 522318417645227. The last choice was '7' and the reinforcement schedule would check the counter of the pair [27], which indicates the frequency of occurrence of this pair in the entire generated sequence up to that point. If the counter of this pair was among the lowest 60% of all possible 64 response pairs, a reward of 1 would result; otherwise, the reward would be 0. In the example, the pair is indeed among the lowest 60% (indeed, the pair had not yet been generated) so a reward of 1 is given. The choice in the previous trial in the example was '2'. This trial would check the counter of the pair [22]. In this case, because the pair occurred before and its choice frequency (or counter) is higher than the other pairs, no reward was given in this trial.

Stable and volatile environments have been studied extensively in reinforcement-learning and decision-making before (Behrens et al., 2007; Browning et al., 2015; Nassar et al., 2010; Simoens et al., 2024). In stable environments, agents learn the high rewarding options and stick to them. In volatile environments, agents are required to explore other options when the reward probabilities shift, i.e., it requires variability in decision-making from time to time, because reward probabilities change regularly across trials. In an adversarial environment, however, the best strategy is to make variable choices all the time. This environment therefore allowed us to study more purely strategically variable behavior. The benefit of simulating performance across all three environments further allowed us to determine how and which parameters need to be adjusted to meet different variability requirements.

We used a parameter optimization strategy to assess which parameter values for ε, α and λ (Eq. 1-3) are optimal in the three environments. Each parameter was optimized in a separate set of simulations. One way to do this, is to use a second RW update rule on a higher level to learn Q-values associated to certain parameter values or intervals (Sikora, 2008). However, this approach

confines these parameters to discrete and arbitrary values or intervals. For this reason, we instead used a policy-gradient update rule, resulting in a continuous parameter space in which the parameters were optimized. The policy-gradient update rule seeks to identify the optimal (most rewarding) values of the parameters within each environment (Sutton & Barto, 2020; van Heeswijk, 2020; Williams, 1992). This method enables us to validate the efficacy of upregulating or downregulating the parameters in the three environments. The policy-gradient method updates the mean and log(standard deviation) of a normal distribution, from which ε, α, or λ were subsequently sampled. The parameter updates occurred on a slower time scale than the value updates (every 10 trials, as in Sikora, 2008).

A first set of simulations optimizes ε. The update rules for the mean and standard deviation of ε can be seen in the following equations:

$$\mu_{\tilde{\varepsilon},t+10} = \mu_{\tilde{\varepsilon},t} + \alpha_\mu r_{b,t+10} \frac{(\tilde{\varepsilon}_t - \mu_{\tilde{\varepsilon},t})}{\sigma^2_{\tilde{\varepsilon},t}} \tag{4}$$

$$\sigma^*_{\tilde{\varepsilon},t+10} = \sigma^*_{\tilde{\varepsilon},t} + \alpha_\sigma r_{b,t+10} \left(\frac{(\tilde{\varepsilon}_t - \mu_{\tilde{\varepsilon},t})^2}{\exp(\sigma^{*2}_{\tilde{\varepsilon},t})} - 1 \right) \tag{5}$$

for $t = 10x$ with $x \in \mathbb{N}$. Here, μ represents the mean of the optimized parameter (in this Equation $\tilde{\varepsilon}$) and σ^* is the natural logarithm of σ, the standard deviation of the meta-learned parameter (i.e., $\sigma^* = \ln(\sigma)$). Both mean and standard deviation have a separate learning rate, respectively α_μ and α_σ. In all simulations, α_μ was set to 0.5 and α_σ was set to 0.1. In the Equations given here, the optimized parameter is $\tilde{\varepsilon}$, which is sampled from a normal distribution, $\tilde{\varepsilon}_t \sim \mathcal{N}(\mu_{\tilde{\varepsilon},t}, \sigma_{\tilde{\varepsilon},t})$. The final ε used in the decision-making of Q-values (Equation 1), is an inversed logit transformation of $\tilde{\varepsilon}$, such that $0 \leq \varepsilon \leq 1$ (i.e., $\varepsilon_t = \frac{\exp(\tilde{\varepsilon}_t)}{1+\exp(\tilde{\varepsilon}_t)}$).

The updates use a baselined reward $r_{b,t}$ (Williams, 1992), by subtracting a weighted reward average from the average of the 10 last received rewards, i.e., $r_{b,t} = \bar{r}_{(t-10):t} - R_{t-11}$, where $\bar{r}_{(t-10):t}$ is the average received reward of the last 10 trials. R_{t-11} is a weighted average of all past rewards from $t = 0$ to $t - 11$, i.e., $R_{t-11} = R_{t-12} + \alpha_R(r_{t-11} - R_{t-12})$, with α_R the reward learning rate, set to 0.25 in all simulations, and $r_{t\text{-}11}$ is the received reward at trial t-11 (the last trial before the last update). The baseline rewards evaluates if the average reward of the last 10 trials that resulted from the last parameter update is better or worse than a weighted average reward of everything that happened before that update, and it uses this information to make the next parameter update.

Similarly, a second set of simulations optimizes the value of $\tilde{\alpha}$, where $\tilde{\alpha}$ is a logit transformation of α (so $0 \leq \alpha \leq 1$). A third set of simulations optimizes $\tilde{\lambda}$, where $\tilde{\lambda}$ is a logit transformation of λ, rescaled in a way that $-3 \leq \lambda \leq 3$. Each set of simulations included 500 runs of 10000 trials in each of the three environments, where each set optimized only one parameter. The other two parameters were fixed to common-sense values. Specifically, the optimization of ε had a fixed learning rate $\alpha = 0.25$ and a fixed unchosen value-bias $\lambda = 0$ (i.e., no choice-bias) in each environment. The optimization for learning rate fixed $\varepsilon = 0.3$ and $\lambda = 0$ in each environment. The optimization for the unchosen value-bias λ fixed $\varepsilon = 0.3$ and $\alpha = 0.25$ in each environment.

In each simulation set, all parameters converged to specific values in each environment. To assess if these values were indeed optimal for each environment, we simulated our model (Equations 1-3) using these values, without the optimization part (Equations 4-5). We calculated the optimal parameter values for ε, α, and λ by taking the average of the logit transformations of the learned means, e.g. for epsilon $\mu_{\tilde{\varepsilon}}$, of the last 100 trials across the 500 simulations. These values were combined with the fixed parameter settings with which the optimization took place. This

resulted in 9 new sets of simulations in each environment with different combinations of parameter values (Table 1). Each of these sets contained 100 simulations over 2500 trials.

Table 1: Parameter settings in simulations

Optimization of ε			**Optimization of α**			**Optimization of λ**		
Stable (sta)	Volatile (vol)	Adversarial (adv)	Stable (sta)	Volatile (vol)	Adversarial (adv)	Stable (sta)	Volatile (vol)	Adversarial (adv)
Opt $\boldsymbol{\varepsilon_{sta}}$	Opt ε_{sta}	Opt ε_{sta}	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3
α = 0.25	α = 0.25	α = 0.25	Opt α_{sta}	Opt α_{sta}	Opt α_{sta}	α = 0.25	α = 0.25	α = 0.25
$\boldsymbol{\lambda}$ = 0	λ = 0	λ = 0	λ = 0	λ = 0	λ = 0	Opt λ_{sta}	Opt λ_{sta}	Opt λ_{sta}
Opt $\boldsymbol{\varepsilon_{vol}}$	Opt ε_{vol}	Opt ε_{vol}	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3
α = 0.25	α = 0.25	α = 0.25	Opt α_{vol}	Opt α_{vol}	Opt α_{vol}	α = 0.25	α = 0.25	α = 0.25
$\boldsymbol{\lambda}$ = 0	λ = 0	λ = 0	λ = 0	λ = 0	λ = 0	Opt λ_{vol}	Opt λ_{vol}	Opt λ_{vol}
Opt $\boldsymbol{\varepsilon_{adv}}$	Opt ε_{adv}	Opt ε_{adv}	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3	ε = 0.3
α = 0.25	α = 0.25	α = 0.25	Opt α_{adv}	Opt α_{adv}	Opt α_{adv}	α = 0.25	α = 0.25	α = 0.25
$\boldsymbol{\lambda}$ = 0	λ = 0	λ = 0	λ = 0	λ = 0	λ = 0	Opt λ_{adv}	Opt λ_{adv}	Opt λ_{adv}

Parameter settings of simulations with model from Equations 1-3. Every simulation set contains one optimized parameter value (opt), the remaining two parameters are set to the values they took in the optimization simulations.

First, we compared the average reward obtained by the model for each simulation setting separately. Second, to validate if the adaptation of these three parameters across contexts was useful, we calculated the average reward obtained by the model across different sets of simulations. Specifically, this included the average reward across the row of each table (optimization of ε, optimization of α and optimization of λ) and across the diagonal of each table. The latter represents a situation where the model adapts one of these parameters across contexts, i.e., it checks if parameter adjustment across environments is useful when responding more or less variable (rather than having fixed parameters across environments).

Additionally, since the goal of the dynamic parameter settings is to adapt to different variability demands, we calculated a variability measure for each simulation set across the diagonal of Table 1. Variability is measured as the second-order entropy, which is a normalized form of the Shannon entropy (Shannon, 1948), corrected for the total number of choice options (Page & Neuringer, 1985). The second-order entropy is calculated as follows:

$$U_2 = \frac{-\left(\sum_k^m (p_{2,k} log_2(p_{2,k})) + U_1\right)}{log_2(m)} \quad (6)$$

in which we used the first-order (and normalized) entropy or U_1-value:

$$U_1 = \frac{-\sum_i^n (p_{1,i} log_2(p_{1,i}))}{log_2(n)}$$

Here, *n* is the total number of possible responses or actions (8 in the case of the simulations) and *k* is the total number of possible response pairs ($8^2 = 64$ in the case of the simulations settings). The frequencies with which each response and each response pair occurs in a generated sequences is given by $p_{1,i}$ and $p_{2,k}$ respectively. A second-order entropy value of 0 means there is no variability in the sequence. A value of 1 means there is maximum variability in the sequence.

Results

First, we simulated the model while optimizing epsilon (ε) according to Equations 4 and 5. Figure 1A shows the average of the resulting sampled ε and the standard errors in each environment. The average of the optimized mean ε of the last 100 trials suggest an optimal ε value of 0.08 in a stable environment, an optimal value of 0.25 in a volatile environment, and an optimal value of 0.86 in an adversarial environment. We took these optimal values and combined them with the fixed α and λ setting (in which ε was optimized, see Table 1 Optimization of ε), to simulate the model from Equations 1-3 in each of the three environments (Figure 1B). Our results show that the highest average rewards in each environment are indeed obtained by using the respective ε value that was optimized in that environment. Moreover, the model obtains a significantly higher reward when ε is adjusted between environments (Figure 1C). Last, by comparing the second-order entropy between the three environments, we demonstrate that the average variability in responding effectively increased across the three environments with increasing ε (Figure 1D).

Second, we performed the same procedure for optimizing learning rate α (substituting α for ε in equations 4-5 and setting the other parameters equal to common sense values). The average

of the optimized mean α of the last 100 trials over all simulations was 0.37, 0.86, and 0.97 in a stable, volatile, and adversarial context, respectively (Figure 2A). As visualized in Figure 2B, using the optimized α values resulted in a slightly better overall reward (for parameter settings, see Table 1 Optimization of α). Similarly, varying α between environments resulted in a slightly higher overall reward, however not significant (Figure 2C). More importantly, Figure 2D confirms that increasing α did enhance response variability where increased variability is needed.

The third and last optimization was for the unchosen value-bias parameter, λ. Here, λ was substituted for ε in Equation 4 and 5. The optimal values for λ were -0.21, 0.24, and 0.83, in the stable, volatile and adversarial context, respectively. This shows that the model learns to valuate λ positively in an adversarial environment, thereby giving unfrequently chosen options more value (i.e., the longer ago an option was chosen, the higher its *Q*-value). In a stable environment, the model learns to devaluate unchosen options, thereby giving the more frequent chosen options more value. The optimal values were plugged into the model of Equations 1-3, together with the fixed ε- and α value (see Table 1, Optimization of λ). Figure 3B shows the average reward obtained with each parameter combination in each environment, clearly giving an advantage of a positive unchosen value-bias in the adversarial environment, and an advantage of a negative unchosen value-bias in the stable environment, but not much effect in the volatile environment. Nonetheless, an environment-specific λ is more beneficial than keeping λ constant (Figure 3C). Last, Figure 3D shows the second-order entropy in each environment, demonstrating that positive λ values lead to a significant increase in variability in an adversarial and volatile environment, as opposed to the λ values in a stable.

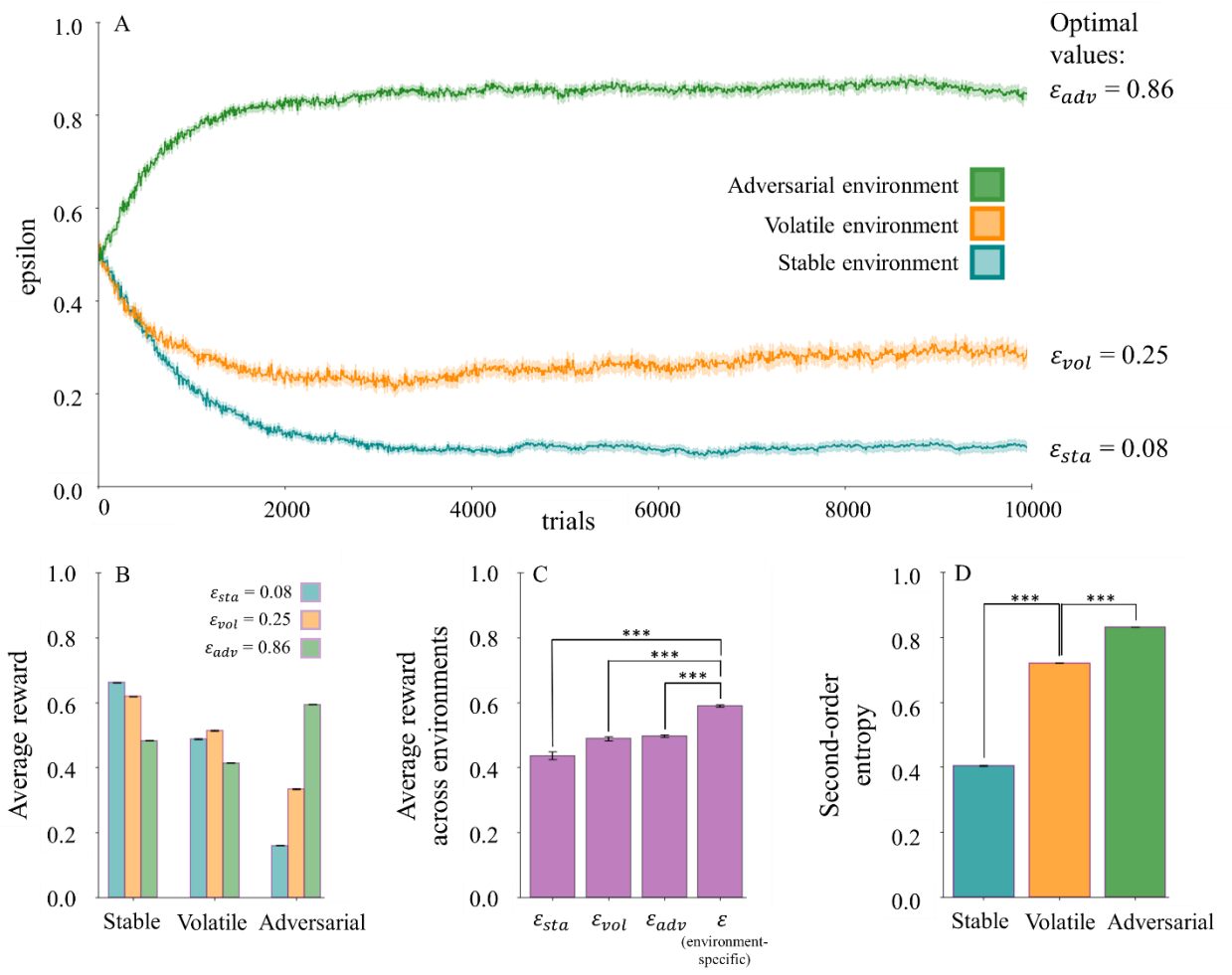

Figure 1: A) Optimization of epsilon (ε). The blue line shows the result for the stable environment, the orange line shows the result for the volatile environment and the green line shows the result for the adversarial environment. The plot shows the average across 500 simulations of the sampled ε, where the shaded areas above and below the lines are standard errors. Learning rate (α) was fixed across environments and equal to 0.25. The unchosen value-bias (λ) was also kept fixed and equal to 0. The optimal values for ε (ε_{sta}, ε_{vol} and ε_{adv}) are the average values of the learned mean ε in the last 100 trials of each simulation. B) Average reward across 100 simulations and 2500 trials obtained by the model from Equations 1-3 using different sets of parameter combinations in the three environments. Each parameter combination contains one of the optimized ε values and the fixed α and λ value that was used in the optimization. C) Average reward across three environments, where the first three bars represent the result when ε is kept constant but equal to one of the three optimal values across environments. The fourth bar shows the result when ε is set to its optimal value in each environment. Rewards obtained by the model are significantly higher when an environment-specific ε is used. D) Average second-order entropy in each environment when using the respective, optimal ε value. Variability significantly increases between the environments going from stable to volatile to adversarial, while using increasing ε.

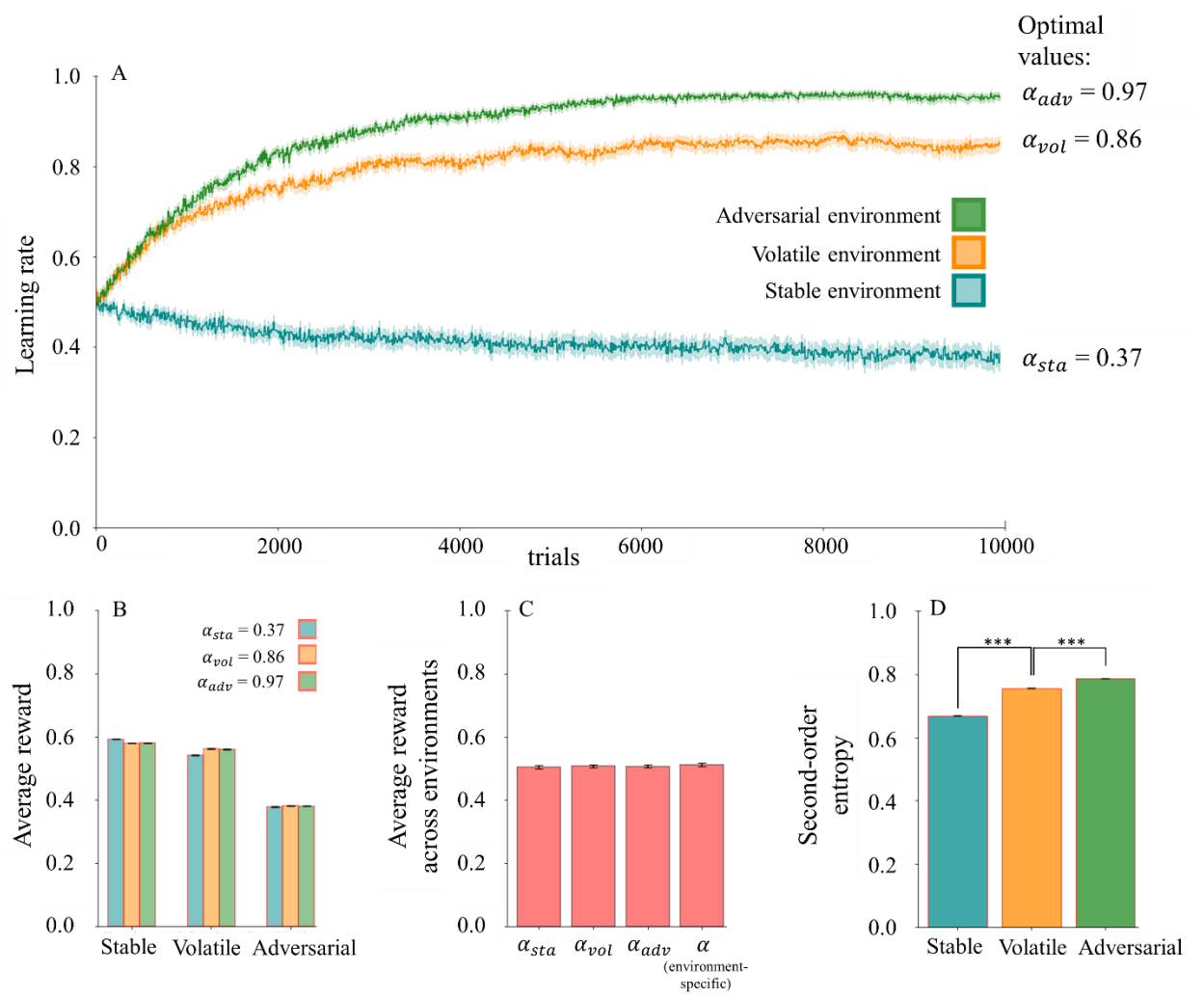

Figure 2: A) Optimization of learning rate (α). The blue line shows the result for the stable environment, the orange line shows the result for the volatile environment and the green line shows the result for the adversarial environment. The plot shows the average across 500 simulations of the sampled α, where the shaded areas above and below the lines are standard errors. Epsilon (ε) was fixed across environments and equal to 0.3. The unchosen value-bias (λ) was also kept fixed and equal to 0. The optimal values for α (α_{sta}, α_{vol} and α_{adv}) are the average values of the learned mean α in the last 100 trials of each simulation. B) Average reward across 100 simulations and 2500 trials obtained by the model from Equations 1-3 using different sets of parameter combinations in the three environments. Each parameter combination contains one of the optimized α values and the fixed ε and λ value that was used in the optimization. C) Average reward across three environments, where the first three bars represent the result when α is kept constant but equal to one of the three optimal values across environments. The fourth bar shows the result when α is allowed to take on its optimal value in each environment. Rewards obtained by the model are higher when an environment-specific α is used, but not significant. D) Average second-order entropy in each environment when using the respective, optimal α value. Variability significantly increases between the environments going from stable to volatile to adversarial, while using increasing α.

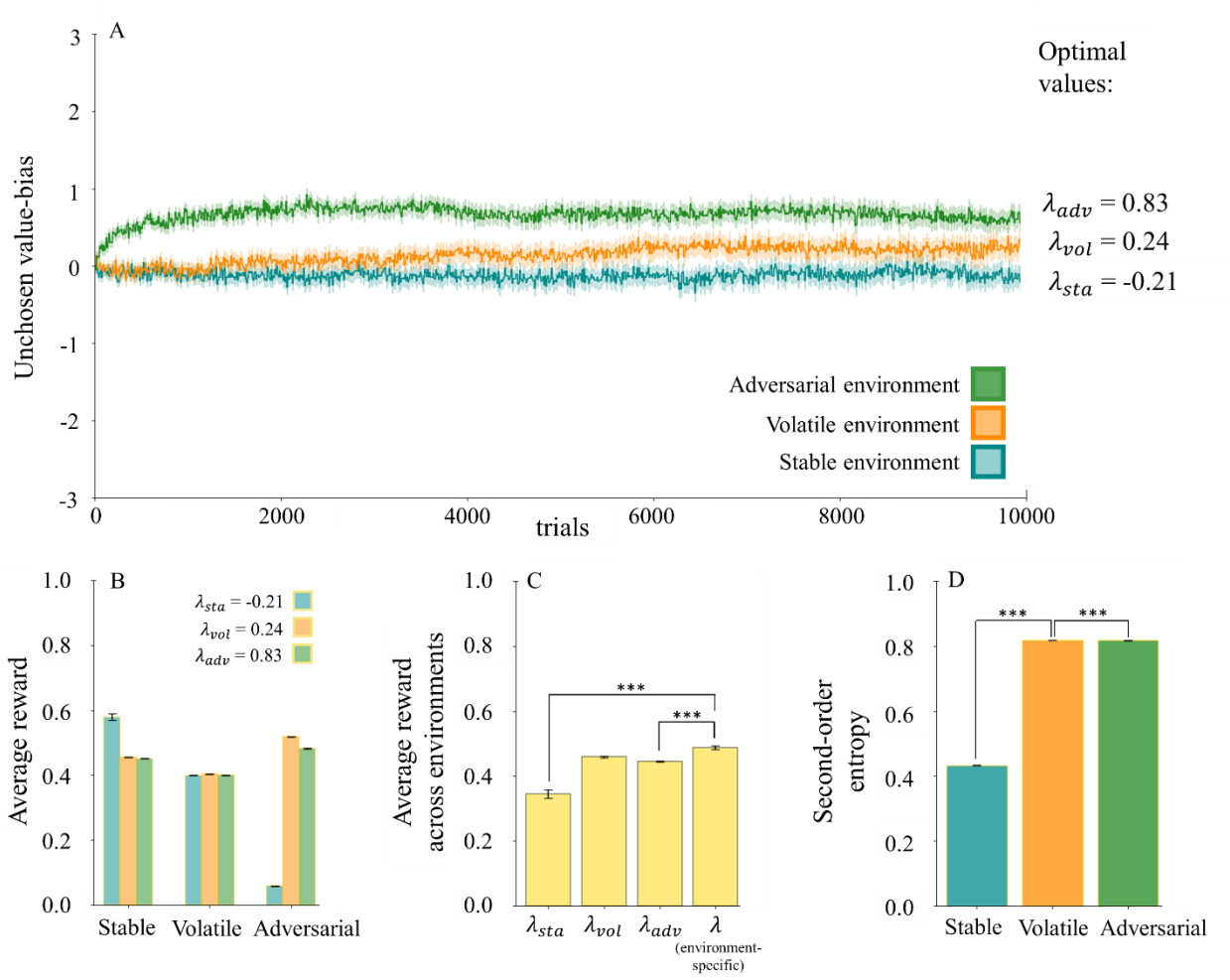

Figure 3: A) Optimization of unchosen value-bias (λ). The blue line shows the result for the stable environment, the orange line shows the result for the volatile environment and the green line shows the result for the adversarial environment. The plot shows the average across 500 simulations of the sampled λ, where the shaded areas above and below the lines are standard errors. Epsilon (ε) was fixed across environments and equal to 0.3. Learning rate (α) was also kept fixed and equal to 0.25. The optimal values for λ (opt λ_{sta}, opt λ_{vol} and opt λ_{adv}) are the average values of the learned mean λ in the last 100 trials of each simulation. B) Average reward across 100 simulations and 2500 trials obtained by the model from Equations 1-3 using different sets of parameter combinations in the three environments. Each parameter combination contains one of the optimized λ values and the fixed ε and α value that was used in the optimization. C) Average reward across three environments, where the first three bars represent the result when λ is kept constant but equal to one of the three optimal values across environments. The fourth bar shows the result when λ is allowed to switch to its optimal value in each environment. Rewards obtained by the model are higher when an environment-specific λ is used. D) Average second-order entropy in each environment when using the respective, optimal λ value. Variability significantly increases between the environments when going from stable to volatile and adversarial, while increasing λ.

Discussion

The aim of Study 1 was to investigate whether we can model and simulate three cognitive mechanisms that have been argued to underlie strategically variable behavior: a stochastic generator (Jensen et al., 2006; Page & Neuringer, 1985), a dynamic reinforcement/extinction process (Machado & Tonneau, 2012; Nergaard & Holth, 2020), and a frequency-based memory (Capone et al., 2014; Oomens et al., 2015). We linked each cognitive mechanism to a computational mechanism, i.e., the dynamics of three parameters in our model. The model consists of a decision-making policy using epsilon-greedy where ε represents the probability to select an action randomly, a standard Rescorla-Wagner update rule for chosen actions with learning rate α, and an additional update rule for unchosen actions involving the addition of an unchosen value-bias λ (Rescorla & Wagner, 1972; Sutton & Barto, 2020). We simulated and optimized this model in three different environments that require an increasing level of variable responding. The goal was for the model to achieve different levels of variability, solely by adapting one of the three parameters linked to an underlying cognitive mechanism. All parameters converged to different values across the different environments, which were lowest for the stable environment and highest for the adversarial environment. Furthermore, these higher values effectively resulted in more variable behavior, as summarized in the second-order entropy. This confirmed that the upregulation of one of these three parameters is sufficient to generate variable choices.

We introduced the parameter λ to stimulate frequency-based memory, allowing the agent to choose more infrequent response options with higher values of λ in order to behave variable. Interestingly, however, we observed negative values for λ in the stable context, which reflects a devaluation of unchosen options: the more an option does not get chosen, the less likely it will be chosen later. In hindsight, this outcome could be expected since, in a stable environment, high

rewarding options are consistent across trials. Once identified, the model should keep selecting those. Positive λ values reflect a valuation of unchosen options: the more an option does not get chosen, the higher its associated Q-value, and the more likely they will be selected (when the greedy option is chosen). This made most sense in the volatile and adversarial environments where reward probabilities often change and exploration is necessary (i.e., volatile environment), or the least frequently chosen options were rewarded and switching to them was beneficial at every trial (i.e., adversarial environment). While the optimal value for λ was estimated higher in the adversarial environment, an estimation of obtained reward rate suggested the optimal value for the volatile environment was also slightly more optimal in the adversarial environment (Figure 3B). It is therefore possible that the policy gradient did not find the optimal parameter and overestimated the value for λ. Although we wanted to remain agnostic about the optimal value and therefore chose to keep the bounds for estimating this parameter relatively wide (from -3 to 3), it is possible that this led the model to explore overoptimistic values of λ that were hard to come back from.

Taken together, our simulations show that adapting any of the three parameters in our model is sufficient to significantly increase behavioral variability. In other words, the three proposed mechanisms each offer a potential way of generating strategically variable behavior in an environment that calls for it, especially the adversarial environment. Building on this, our next question was to study which mechanism humans, and non-human animals, effectively employ. In our next study, we thus set out to estimate the origin of variability in humans, pigeons and rats, who were exposed to an adversarial environment. By fitting our model to observed behavior, the resulting parameter values could reveal insights into strategically variable behavior across these different species. From Study 1, we predict that contributions of one (or more) of the mechanisms would be reflected by relatively high values for the associated parameter estimates.

Study 2: Model estimation of human, pigeon, and rat behavior in adversarial environments

Next, as a first empirical application of our model, we estimated the model parameters in three existing datasets of humans, pigeons, and rats who were exposed to an adversarial environment. In this environment, our model predicts higher values for one (or multiple) of the three critical parameters (ε, α, and λ), in line with the three proposed underlying cognitive mechanisms. Importantly, we do not expect these cognitive mechanisms to be mutually exclusive. A mix of two or even all three mechanisms is plausible. We had no strong predictions as to which mechanism, or combination thereof, human and non-human animals would use to generate strategically variable behavior but previous studies have investigated the effect of (variants of) the critical parameters in our model, although never in an adversarial environment.

For example, it has shown that in a more volatile environment, humans tend to use both higher learning rates (α) and more randomness (ε) in their decisions (Behrens et al., 2007; Goris et al., 2021; Simoens et al., 2024; Verbeke & Verguts, 2024), which suggests they may also be able to flexibly upregulate those parameters when faced with adversarial environments requiring yet more variability. Similarly, Jin et al. (2024) showed that pigeons can also dynamically adjust their learning rate (α) and the randomness (ε) that goes in their decisions while learning a task. Rats have similarly been shown to dynamically adjust learning rates (α) over time, and depending on the uncertainty of the choice options (Funamizu et al., 2012). Together, these studies suggest that all three species may be able to also find optimal (relatively high) ε and α values specific for an adversarial context.

The unchosen value-bias (λ) or variants thereof have also been used in other studies before. Often referred to as 'stickiness' or choice persistence (but simultaneously allowing for a switching bias), this parameter has been shown to be a valuable addition to RW models in probabilistic reversal learning tasks (comparable to a volatile environment) with two choice options in humans (Eckstein et al., 2022). More generally, Palminteri et al. (2016) showed that, in a probabilistic learning task, adult human participants use a counterfactual updating rule (similar to the unchosen value-bias but not exactly the same), in addition to the classical Rescorla-Wagner updating rule for chosen options. On the contrary, adolescents did not seem to benefit from a counterfactual updating rule.

Similarly, Laurent and Balleine, (2015) suggest that also rats seem to encode unchosen action values, and may use counterfactual reasoning in their decision-making. Counterfactual (or unchosen option) value updating has not been studied or observed in pigeons before. Moreover, it remains contested whether non-human animals have similar action and event memory as humans do (Lind & Jon-And, 2024), questioning their ability to flexibly use frequency-based memory in an adversarial environment.

More broadly focusing on the considered mechanisms rather than specific parameters, Machado (1993) suggested that some animals may rely on both memory strategies and a stochastic generator in an adversarial environment, further depending on the demands of the variability contingency. De Souza Barba (2015) argued that unpredictable responding is a default setting in humans and nonhuman animals when no actions are consistently rewarded, aligning with a stochastic generator. Although this may be the case, it remains somewhat unclear how much people can really tap into this stochastic generator, even when explicitly instructed to. For instance, in RNG tasks, it is a consistent observation that humans show certain biases when trying to be

random (Capone et al., 2014; Oomens et al., 2015; Ross, 1955; Wagenaar, 1972). These biases include repetitions (repeating the same digits), cycling (repeating a certain sequence of digits) and seriation (changing digits by adding or subtracting one or two units) (Ginsburg & Karpiuk, 1994). Such systematic processes seem more consistent with a memory-based mechanism. It has also been shown that RNG tasks recruit more attentional resources in comparison to ordered number generation tasks (Joppich et al., 2004). These biases and the allocation of attention suggest that humans might rely more on memory when trying to be variable.

Methods

Datasets

The human and pigeon dataset were first described in Jensen et al. (2006). The objective in this study was to investigate whether the level of variability differed when subjects were exposed to an adversarial environment with two, four, or eight choice options. The pigeon dataset included five subjects, who were maintained at 85% of their normal weight and performed around 150000 trials each. The pigeons had to peck on a screen on different response squares. All five pigeons were exposed to three different conditions (with two, four or eight response options), of which the response square orientations are visualized in Figure 4A (only the squares were visible to the pigeons).

An example of a trial for the pigeons can be seen in Figure 4B. A trial started with a fixation square. When pigeons pecked this square, the response options (2, 4, or 8) were projected. Pigeons had to peck one of these response squares; if their choice satisfied the variability contingency, a cue for reinforcement was shown (a green asterisk), otherwise the initial fixation square was projected again, initiating a new trial. If the variability contingency was met and the green asterisk appeared, pigeons had to peck it, which led in 25% of all trials to an access to food for 1.2s. After

this (and in the other 75% of the cases), the initial fixation square was shown again, and a new trial started. Pigeons performed approximately 75 sessions of 2000 trials each, where the sessions cycled between a fixed order of the two-choice condition, four-choice condition and eight-choice condition. Only the final 18 sessions were used for analysis (6 sessions per condition), resulting in 12000 trials per condition and per pigeon.

Similar to the reinforcement schedule used for the simulations of the adversarial environment in Study 1, the variability contingency required a response to complete a sequence that was among the least frequent emitted sequences. To keep the number of possible sequences constant in each condition, the two-choice condition kept track of frequencies of six consecutive responses ($2^6 = 64$), the four-choice condition kept track of three consecutive responses ($4^3 = 64$), and the eight-choice condition kept track of two consecutive responses ($8^2 = 64$; as explained in Study 1). This resulted in 64 possible sequences in each condition. For example, if in the two-choice condition a sequence '10110100101000101011110101O' was pecked, the last pecked option was 0, which completed the 6-response sequence '101010'. The second last pecked option was 1, which completed the sequence '110101'. For each of these response sequences, the frequency was tracked using a counter. An example of the 8 choice condition can be found in Study 1.

All 64 counters in all three conditions were initialized to a value of 20 units. Each response increased the value of the counter of the sequence it completed by one unit. To maintain a constant sum across counters, 1/63 was subtracted from all other 63 frequencies. The variability contingency was met when the counter of the completed sequence was below 21.6. When a reward was given, all counters were multiplied by an amnesia coefficient of 0.984, making it more likely

that frequencies of less recent sequences were decreased sufficiently to meet the variability contingency.

The human dataset was similar to the design with pigeons and is also described in Jensen et al. (2006). Six college students participated and were rewarded $40 and the possibility to earn $25 if they generated sequences that were *more* variable (higher second-order entropy) than the least variable sequence generated by the pigeons (i.e., their second-order entropy needed to be higher than the lowest pigeon second-order entropy), and $50 if they could produce the most variable sequence (highest second-order entropy overall). Instead of pecking, the human participants responded on a numeric keypad of an e-Mac computer. The available keys were visually represented on screen where the key's image lighted up after responding. The experiment differed from the pigeon experiment in that center-key presses and a response to a green asterisk were not necessary. Moreover, human participants were rewarded with points instead of food. Each participant performed blocks of 150 trials in a randomized order but ensuring that each set of three consecutive blocks contained the two-, four-, and eight condition at least once. The participants eventually performed a variable number of trials in each condition. Total number of trials ranged in the two-choice condition between 3770 and 35960 trials, with an average of 17202 trials. In the four-choice condition, the total number of trials ranged from 3915 to 35960 trials, with an average of 17182 trials. In the eight-choice condition, the total number of trials ranged from 4060 to 35815 trials, with an average of 17255 trials.

The last dataset contained twenty rats, and has been made public online (Jensen, 2018, https://osf.io/5h4zx/). The rats were put in a cage that had 5 different response operands (3 keys and 2 levers, see Figure 4C). Responses had to satisfy a variability contingency (specified shortly) in order to result in a reward. After every response, a brief 2.1 kHz tone was presented during an

interresponse time (IRT) of 0.5s. During this time, response operand lights were darkened; responses during the IRT were not accounted for and reset the IRT. The response-generated tone was followed by a sequence of decreasing tones (2.6 kHz, 2.85 kHz, 3.1 kHz) lasting 1 s, after which a 45 mg food pellet (the reward) was presented. Besides responses to the operands during the IRT or during food delivery, all responses were effective. The rats had five possible responses: right key, center key, left key, left lever and right lever.

All twenty rats were exposed to eight different experimental conditions, among one was close to the above-described adversarial environment. For our purposes, we only used the data from this adversarial environment (the V4 schedule in the original preprint Jensen, 2018). Ten rats performed 20 sessions of this adversarial schedule across two phases in the experiment, while another ten rats performed 10 sessions of this schedule. One session lasted 90 minutes and rats performed between 215-1606 trials per session (on average 900 trials), where each trial was one response.

Similar to the reinforcement schedules described above, the frequencies of response triplets were tracked with counters, being $5^3 = 125$ different counters, initialized at 1. On each trial, one response was made and the *relative* counter of the completed response triplet was checked (i.e., consisting of the current response, and the last two responses, using a moving window for each new trial). If the relative counter (the counter divided by the sum of all counters) was lower than a threshold of 0.0112, the variability contingency was met and food would be presented. After every trial, the counter of the completed response triplet was increased by one unit and an amnesia coefficient (0.95) was applied to all counters to give more weight to triplets that were less recently emitted.

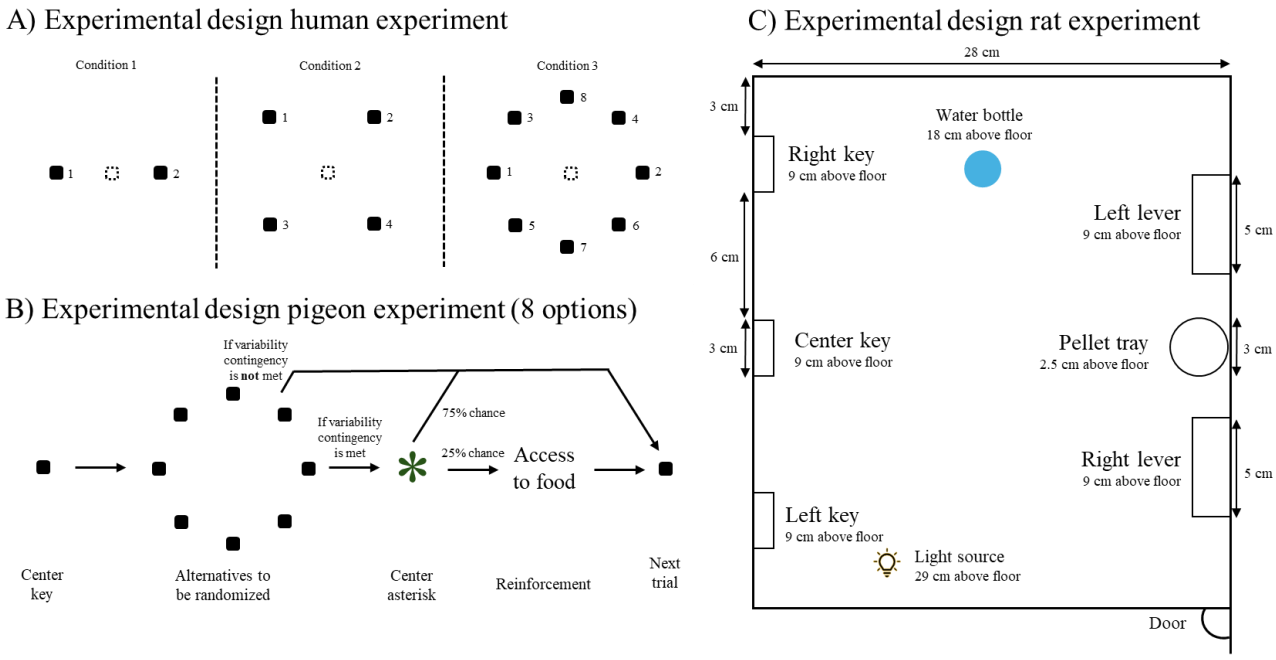

Figure 4: A) Example of the response square options pigeons (and humans) saw. Left: condition with two options. Middle: condition with four response options. Right: condition with eight response options. The numbers were not visible to the participants. Source: Jensen, et al (2006). B) Diagram of a trial in the experiment with pigeons (Jensen, 2006). C) Cage in which the rats were placed for the experiment (Jensen, 2018).

Model simulations, recoveries and model fits

First, because the experimental setups and variability contingencies between datasets were slightly different, we simulated each version of the adversarial context (human/pigeon datasets with 2, 4 and 8 choice alternatives and the rat dataset with 5 alternatives) using the model and the policy-gradient parameter optimization. We compared these simulations with our own adversarial environment from Study 1 to look for differences in optimal parameter values between the adversarial environments. The main difference between our earlier simulations and the human/pigeon datasets was that in the experimental data, frequencies of emitted response sequences were tracked with absolute counters, and a fixed threshold was used to check if these counters were low enough in order to meet the variability contingency. In the rat dataset,

frequencies of response sequences were calculated as relative counters (i.e., by dividing the absolute counter of a response sequences by the total amount of trials). These relative counters needed to be lower than a threshold to meet the variability contingency. In our adversarial environment from Study 1, absolute counters were used to track frequencies of response sequences, but the threshold was relative, i.e., the counters needed to be in the lowest 60% in order to be rewarded. We simulated each variation of the environment using Equations 1-3, and simultaneously optimized epsilon, learning rate or unchosen value-bias (lambda) using Equation 4-5, similarly as in Study 1. Each adversarial environment (5 in total) was simulated for each optimization 100 times over 10000 trials. Updates of the parameters happened every 10 trials.

Next, we fitted our model on the three animal datasets and estimated the three critical parameters: epsilon, learning rate and unchosen value-bias (the ε-α-λ model, Equations 1-3). This is our main model from which we compared the parameter values, where high parameter values could be linked to the associated mechanisms underlying variable behavior. Additionally, we fitted an alternative model to evaluate the influence of adding λ. In the second model, we set the unchosen value-bias to 0, in order to check whether adding this parameter (in the first model) was a valuable addition to a regular Rescorla-Wagner model (i.e., the ε-α model, Equation 1-2).

We performed a model recovery with these two models (Wilson & Collins, 2019). For each model, 1000 datasets were simulated for 500 trials in the 4 adversarial environments resembling the human/pigeon (2, 4 and 8 choice alternatives) experiment and the rat experiment. For each simulated dataset, parameters were sampled from uniform distributions (ε and α between 0 and 1; λ between -3 and 3). The 2000 resulting simulated datasets in each environment were used to fit the two models by minimizing the negative loglikelihood of each model. Best model fits were evaluated using the Bayesian Information Criteria (BIC) values. This resulted in a confusion

matrix, showing the probability of the model being identified as the best fit when the data was actually simulated by that model.

For the winning model, we additionally performed a parameter recovery. To do this, we simulated 500 datasets of 500 trials from the model in the 4 adversarial environments, with randomly sampled parameters (from the same distributions as just mentioned). Then, we minimized the negative loglikelihood of the function from which the data was simulated and estimated the parameters. Correlations between true and estimated parameters were calculated using R^2 value and Pearson- and Spearman correlations.

Last, we fitted the two models on the three datasets (human, pigeon, rat) using negative loglikelihood minimization. BIC values were used to assess which model fitted the data of each subject best. For the ε-α-λ model, we further estimated the respective parameters for each subject in each dataset. Because the 20 rats performed multiple sessions of the experiment on multiple days, resulting in multiple datasets for the same rat subject, we estimated the parameters of the model for each session/dataset separately and took the average of the estimates for each rat.

Results

Model simulations in experimental conditions

First, we simulated the five different adversarial environments with our model and the policy-gradient algorithm (Equations 4 and 5), optimizing either epsilon (ε), learning rate (α) or unchosen value-bias (λ), to evaluate whether the small differences between adversarial environmental setups resulted in a difference in optimal parameter values. In line with Study 1, Figure 5A shows that the optimization of ε led to relatively high optimal values, in all but one of the five adversarial environments. This shows that the number of alternatives has an effect on

optimal ε values, with lower ε values when there are less response alternatives. Additionally, the usage of relative frequency counters or a relative threshold to check the counters (such as in the rat- and Study 1 environment), also led to overall higher optimal ε values. Similarly, Figure 5B shows the optimization of α in the five different adversarial environments. The higher α in these last two environments may also be attributed to the usage of relative counters/thresholds. That is, larger α seemed especially beneficial in those environments, but not necessarily in the environments used in the human/pigeon study. Last, Figure 5C shows the optimization of λ. These simulations consistently showed that positive values of around 1 were preferred for lambda.

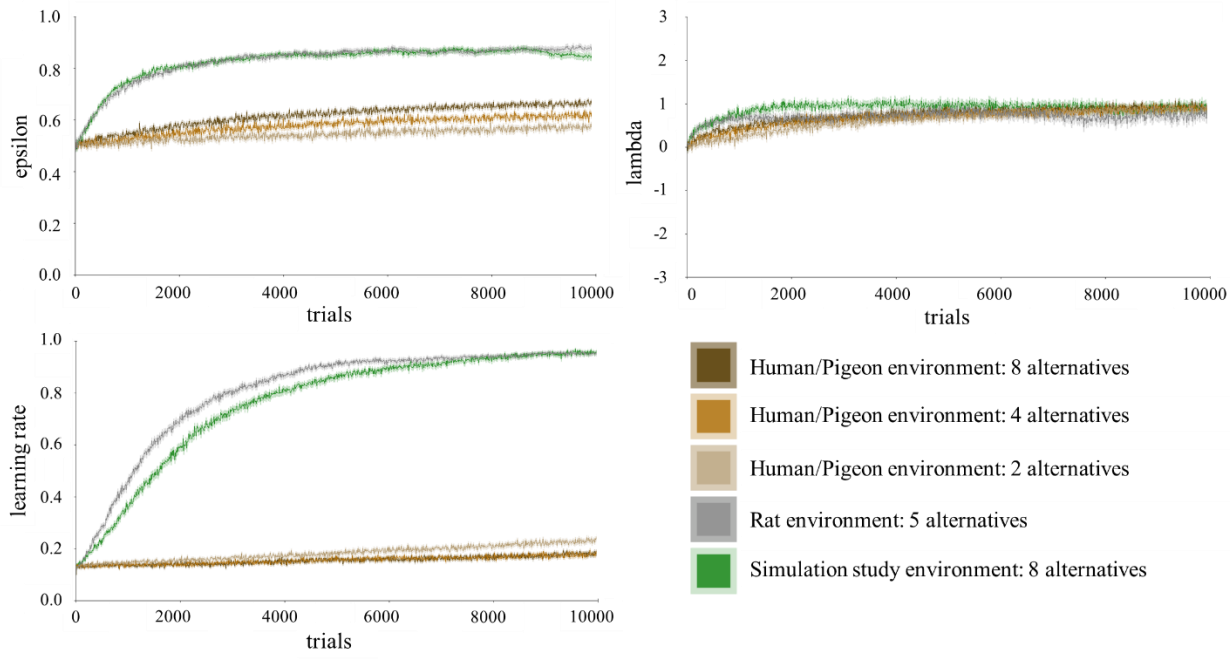

Figure 5: Simulation of five different adversarial environments: the 2 alternative (dark brown)-, 4 alternative (light brown)-, and 8 alternative (beige) experimental setups from the human- and pigeon experiments from Jensen and colleagues (2006), the 5 alternative condition (grey) of the rat experiment from Jensen (2018), and the simulated adversarial environment from Study 1 (green) in this paper. Darker lines are average sampled parameters (from the learned mean and standard deviation), shaded areas are standard errors. All simulations were done with the model in Equations 1-3. The optimization of parameters was done using Equation 4 and 5, where for the optimization of learning rate ε is substituted for α, and in the optimization of lambda $\tilde{\varepsilon}$ was substituted for λ. A) Optimization of epsilon. B) Optimization of learning rate. C) Optimization of lambda.

Model recovery and parameter recovery

Before fitting the ε-α-λ model and the ε-α model to the data, we also performed a model recovery in the adversarial environments resembling each of the 4 experiments. The confusion matrices can be seen in Figure 6. It shows that the ε-α-λ model and ε-α model could be identified well above chance level.

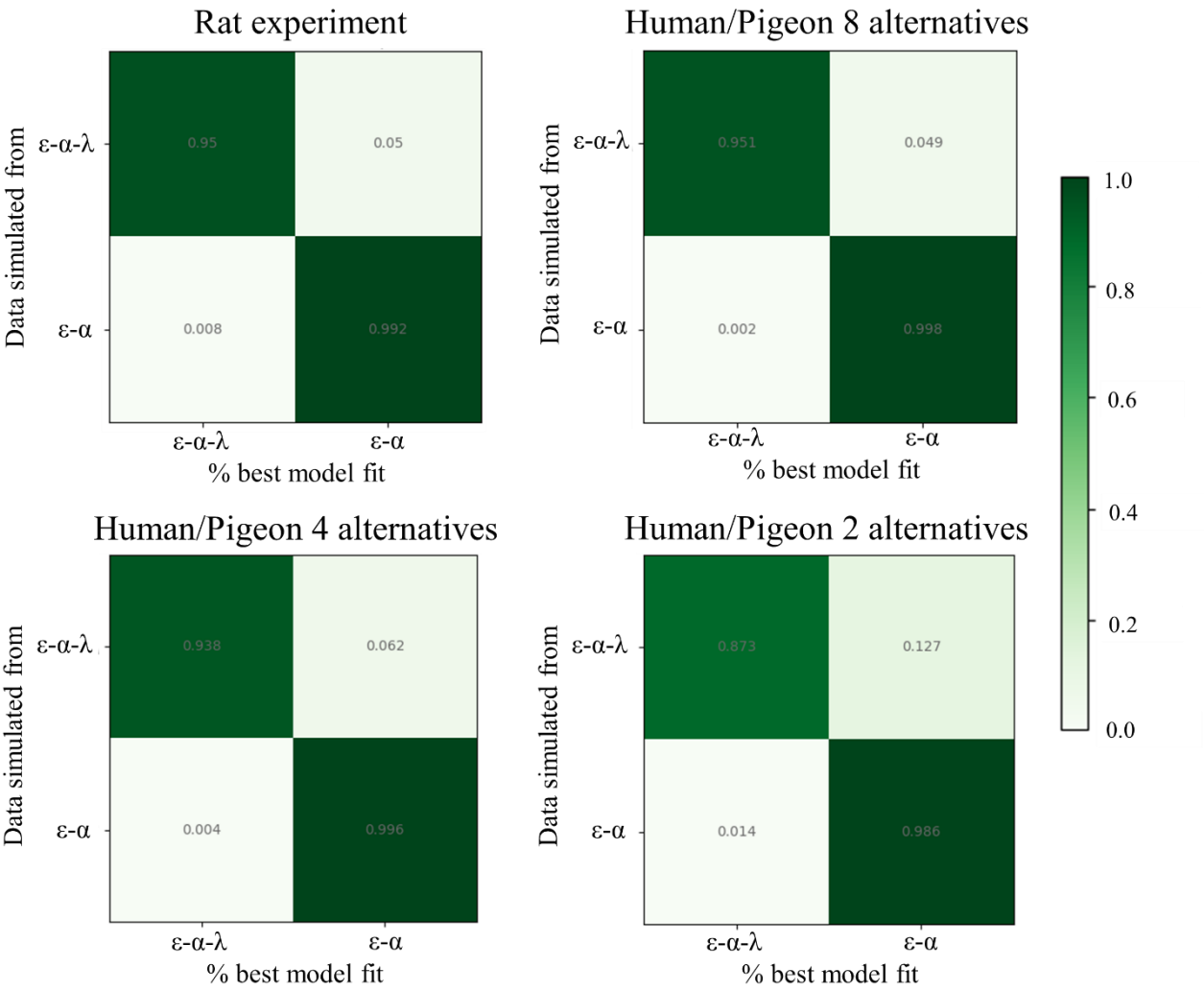

Figure 6: Confusion matrix with two models: the ε-α-λ model, the ε-α model. 1000 datasets were simulated for each model over 500 trials, according to the rat experiment environment and the human/pigeon experimental environment with 8, 4 and 2 choice alternatives. All these simulated datasets were used to fit each of the two models using a negative loglikelihood minimization. The confusion matrix gives the probability that the model on the x-axis result in the best model fit (based on BIC values), given that the data was simulated from the model on the y-axis.

We also performed a parameter recovery analysis for the ε-α-λ model in all adversarial environments resembling the 4 experiments. The true vs recovered parameters for simulations of the rat experiment can be seen in Figure 7. The recoverability of epsilon and learning rate was high (R^2 = 0.99 and 0.80 respectively), meaning that the model estimation process is accurate in finding the values that generated the data. The recovery of λ is accurate within the [-1,1] bounds, but not very recoverable outside of these bounds. This is possibly because values outside these bounds surpassed the value of our reward signal (i.e., 1).

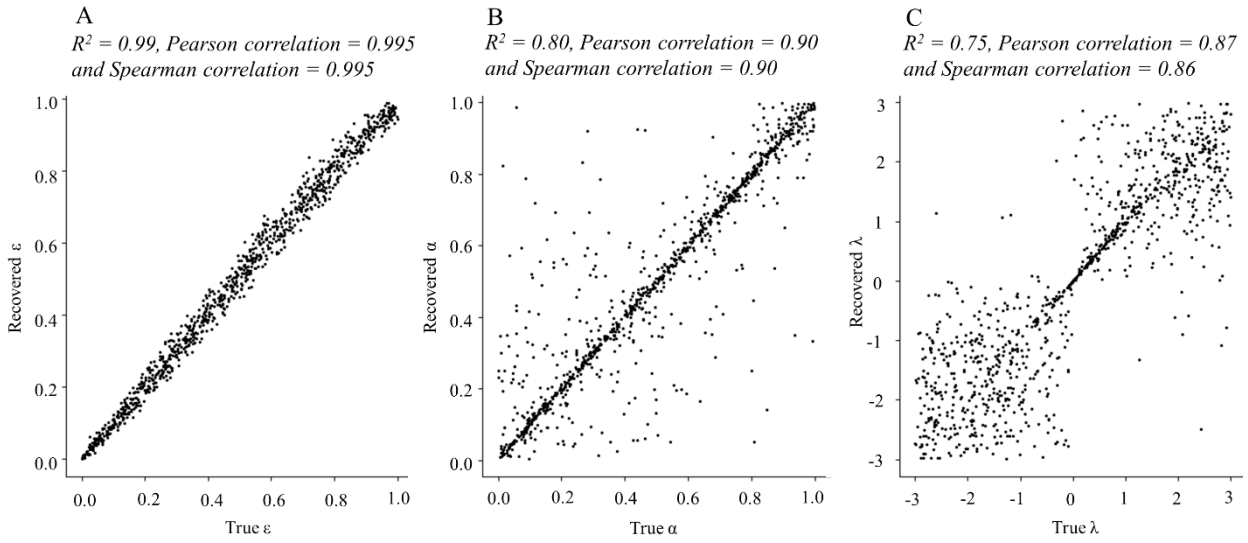

Figure 7: Parameter recovery of the ε-α-λ model simulated in the rat experimental environment. A) Recovery of ε in the ε-α-λ model (R^2 = 0.99, Pearson correlation = 0.995 and Spearman correlation = 0.995). B) Recovery of α in the ε-α-λ model (R^2 = 0.80, Pearson correlation = 0.90 and Spearman correlation = 0.90). C) Recovery of λ in the ε-α-λ model (R^2 = 0.75, Pearson correlation = 0.8 and Spearman correlation = 0.86). All correlations where significant (p < 0.01). We also did parameter recoveries in the human/pigeon experimental environments, with 2, 4 and 8 choice alternatives.

In the 8 choice environment, recovery of ε in the ε-α-λ model achieved an R2 = 0.98, Pearson correlation = 0.992 and Spearman correlation = 0.992, recovery of α achieved an R2 = 0.81, Pearson correlation = 0.90 and Spearman correlation = 0.90 and recovery of λ achieved an R2 = 0.58, Pearson correlation = 0.78 and Spearman correlation = 0.77. In the 4 choice

environment, recovery of ε in the ε-α-λ model achieved an R2 = 0.99, Pearson correlation = 0.995 and Spearman correlation = 0.995, recovery of α achieved an R2 = 0.83, Pearson correlation = 0.91 and Spearman correlation = 0.91 and recovery of λ achieved an R2 = 0.57, Pearson correlation = 0.77 and Spearman correlation = 0.77. In the 2 choice environment, recovery of ε in the ε-α-λ model achieved an R2 = 0.97, Pearson correlation = 0.987 and Spearman correlation = 0.986, recovery of α achieved an R2 = 0.68, Pearson correlation = 0.84 and Spearman correlation = 0.84 and recovery of λ achieved an R2 = 0.59, Pearson correlation = 0.78 and Spearman correlation = 0.79.

Model fits

Finally, we looked at the human-, pigeon-, and rat datasets. First, we examined the variability within the subjects' generated sequences by calculating the second-order entropy per subject (Equation 6) to establish whether all species generated variable behavior. Because subjects performed a varying number of trials, we first calculated the second-order entropy based on the last 500 trials, to keep the information constant between all subjects. The average second-order entropy per species and for the different choice conditions can be seen in Figure 8. The average second-order entropy among human subjects was 0.5 (2 alternatives), 0.74 (4 alternatives) and 0.79 (8 alternatives). The average second-order entropy for pigeons was 0.5 (2 alternatives), 0.73 (4 alternatives) and 0.79 (8 alternatives). Finally, for rats, the average second-order entropy (5 alternatives) was 0.74. Our simulations of Study 1 consistently showed second-order entropy scores around 0.8 in an adversarial environment (Figure 1D, 2D and 3D), that were significantly higher than the second-order entropy scores in a stable and volatile environment (with the exception of the λ optimization simulations, where second-order entropy scores in a volatile vs adversarial environment were similar). The variability levels generated by humans, pigeons and

rats as seen here are comparable to adversarial second-order entropy scores in our simulations (with the exception of the relatively low second-order entropy in the 2 choice alternatives contexts), confirming that these subjects indeed exhibited strategically variable behavior.

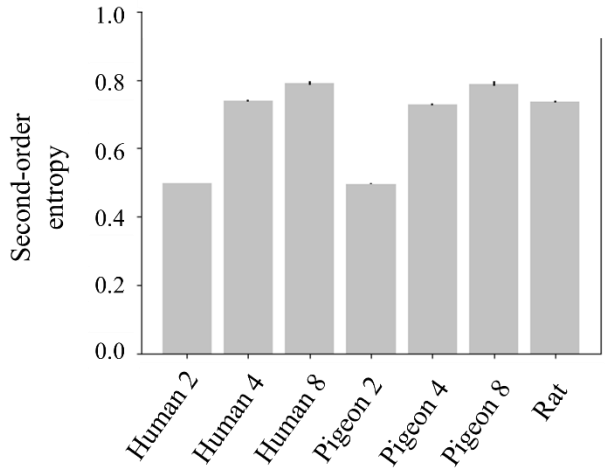

Figure 8: Average second-order entropy for each species' dataset (last 500 trials). A second-order entropy of 0 means there is no variability within the sequence. A second-order entropy of 1 indicates that all possible pairs of alternative options occur equally often in the sequences.

Next, we fitted our two models. The ε-α-λ model gave the best fit for all subjects of all species in all conditions, except for two human subjects where the ε-α model fitted once better in the 2 choice condition and once better in the 4 choice condition. Table 2 provides the average BIC values for each model.

Table 2: Average BIC values of each model for each dataset (human, pigeon and rat).

	ε-α-λ model	**ε-α model**
Human	47240	47758
Pigeon	33042	33286
Rat	2862	2935

Average BIC values for each model and each species' dataset (a lower BIC value means better fit). Each subject/condition was fitted on the two models. The ε-α-λ model fits best to all datasets.

Next, we studied the parameter values. Figure 9A shows that for all species, ε estimates were high. The average value of ε in the ε-α-λ model for the humans was 0.92 (range: 0.77-0.99), for the pigeons also 0.92 (range: 0.88-0.95), and for the rats 0.85 (range average per rat over days: 0.76-0.92).

The α estimates (Figure 9B) showed a less consistent pattern, but more of a broad range of individual differences for humans and pigeons. Interestingly, this result is in accordance with the model simulations suggesting that the optimal value was not necessarily a high α in the precise variability reward contingencies that humans and pigeons were subjected to. Accordingly, the α estimates ranged from 0.17 to 1 with an average of 0.76 for humans, and from 0.02 to 0.91 with an average of 0.4 for pigeons. In the rat dataset, however, α estimates were higher, with an average of 0.88 (range 0.66-0.98). This is again consistent with our simulations showing that the specific variability reward contingency used for rats did benefit from higher learning rates.

Finally, we turned to the λ estimates to investigate how much different species up- or down-valued unchosen actions, with an upregulation shown to be beneficial in generating variable behavior and obtaining more reward in these adversarial environments. Interestingly, λ estimated (Figure 9C) showed species-specific differences. Specifically, λ values were positive for all human subjects and sessions, except for two, with an average of 1.29 (range: -2.09 – 2.69) in the ε-α-λ model. This same model showed an average λ of -1.32 (range: -2.46 – -0.16) for pigeons and -0.84 (range: -1.68 – -0.2) for rats. This suggests that only human subjects show positively valued choice-bias estimates, meaning that only human subjects seemed to upvalue unchosen options. Pigeons and rats, in contrast, show a devaluation of unchosen options, even though this is not optimal in an adversarial environment (Figure 5C).

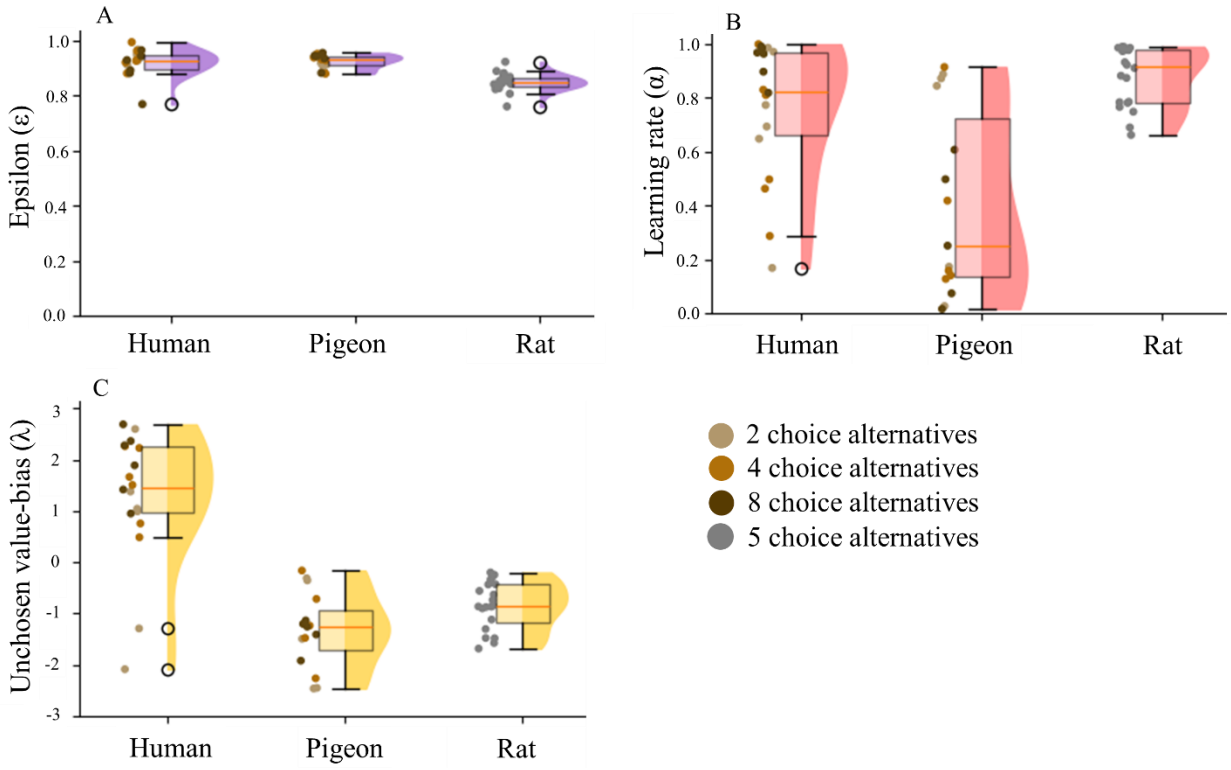

Figure 9; Parameter estimates of the ε-α-λ model. Every dot is a parameter estimation for one dataset belonging to one subject. Color of dots represent the number of choice alternatives that were available to the subjects in each experiment, where beige represent the condition with 2 alternatives, light brown represents the condition with 4 alternatives, dark brown represents the condition with 8 alternatives and grey represents the condition with 5 alternatives. A) Estimates of epsilon in the ε-α-λ model. B) Estimates of learning rates in the ε-α-λ model. C) Estimates of lambda in the ε-α-λ model. In each plot, data is grouped per species. For the rats, every dot is the average of the parameter estimates for that rat across all days.

Discussion

In this study, we examined empirical data from humans, pigeons and rats who were exposed to an adversarial environment. Although the experimental settings differed slightly between these studies, subjects were always rewarded more if their currently chosen option completed a sequence that was among the previously least frequent emitted sequences. These adversarial environments were designed in such a way that highly variable responding is the most optimal strategy in order to achieve a high reward rate, which was also observed as indicated by the average second-order

entropy values, especially in the four-, and eight-choice conditions for humans and pigeons, and for rats.

We fitted the ε-α-λ model to the empirical datasets, and further compared it to a model lacking λ. For most subjects the ε-α-λ model fitted best. In all models, and across species, we observed high ε values, consistent with the idea of a stochastic generator. The learning rate spanned almost the entire range of available values for humans and pigeons, but were high for rats in the ε-α-λ model ($\alpha > 0.66$). This large range of learning rates for humans and pigeons was potentially a consequence of the specific reinforcement schedule as also suggested by our simulations.

Interestingly, λ was almost exclusively positive for humans, and negative for pigeons and rats. This suggests that only humans upvalue unchosen options in an environment that requires them to respond more variably. That is, while most of human decision-making is consistent with a stochastic generator (as seen in the high epsilon values), the proportion of times where decision-making relied on learning (Q-values), they exhibited behavior that seemed fitting to upvalued actions that had been chosen less. This same unchosen value-bias was negative for pigeons and rats. Although for pigeons and rats, the decision-making is also best captured by a stochastic process (high epsilons), negative unchosen value-biases suggest that in those instances that they were relying on learned values, they showed more of a repetition bias towards the previously made choice.

General Discussion

We linked three potential cognitive mechanisms for generating strategically variable behavior to a computational framework. Specifically, we evaluated a model that included a stochastic generator that assumes the ability to respond randomly by blocking out past information

(Page & Neuringer, 1985), a dynamic reinforcement and extinction process that relies on learning from past outcomes (Machado, 1992; Nergaard & Holth, 2020), and a frequency-based memory that keeps track of past action frequencies. We argued that each mechanism can be described by the regulation of a specific parameter in the computational framework and that by adjusting this parameter, different levels of strategically variable behavior can be generated. In Study 1, we showed that upregulating one of the three parameters is sufficient to obtain more strategically variable behavior. In Study 2, we fitted our model to empirical datasets with humans, pigeons, and rats, and found that all species show parameter values compatible with a stochastic generator. Crucially, however, only humans additionally show an upvaluation of unchosen options consistent with a frequency-based memory, and only rats show consistently high learning rates.

To investigate the internal validity of our model, we first simulated the model in three environments (stable, volatile and adversarial) and used a policy-gradient method to optimize each parameter linked to a cognitive mechanism. Specifically, the stochastic generator mechanism predicted an upregulation of epsilon when going from a stable to a volatile and finally to an adversarial environment. When ε was free to be learned, it indeed resulted in the upregulation of ε across these environments, successfully resulting in a maximum average reward and an increase in variability across the environments in the predicted direction. The dynamics reinforcement/extinction learning mechanism predicted the upregulation of learning rate α across these environments. Optimization simulations with a policy-gradient showed that α indeed adjusts in the expected directions in each environment. However, the effect of an upregulated α on performance and variability measures was smaller than an upregulation of ε. Last, the frequency-based memory mechanism predicted the upregulation of λ with increasing variability demands. Here too, the simulations showed that this unchosen value-bias indeed increased across

environments. That is, adjusting λ leads to different levels of variability in decision-making. This shows that each of these three mechanisms can indeed establish different variability demands in different environments.

Optimizing parameters while the model interacts with the environment is an instance of meta-learning, as often applied in reinforcement learning and machine learning (Hutter et al., 2019). Meta-learning algorithms (such as Eq. 4 and 5) optimize parameters on a continuous scale while the learning rule and decision rule (such as Eq. 1-3) carries out the task (Raza Ali et al., 2020; Sikora, 2008). A similar approach was used by Sikora (2008) to meta-learn (or optimize) the temperature parameter (comparable to epsilon) in a Softmax decision policy. Instead of a policy gradient method, they used an RW learning rule, to learn the values of different intervals from which the temperature parameter was sampled. Their meta-learning algorithm quickly learned the optimal temperature value in a stable environment and returned better rewards with increasing volatility of the environment. Yet another way to update parameters of reinforcement learning models is with the Stochastic Real Value Units algorithm, as used by Schweighofer & Doya (2003). This algorithm learns meta-parameters using a stochastic gradient. Similarly as with the policy-gradient of Study 1, a mean value is learned based on a reward feedback signal but the noise term in this case is random (and not learned as with the policy-gradient). Their algorithm also quickly learned the optimal learning rate, temperature and discount factor in a Markov decision problem task, making the algorithm adaptable to the settings of the task environment. The goal of these meta-learning algorithms was similar to our study: optimizing parameters to fit the environment in which the decision is made. Meta-learning parameters in real-time doesn't confine these parameters to single values and allows for a more flexible response to changes in the environment. Intuitively, this concept more closely aligns with human and non-human animal

behavior, as they also need to be capable of adapting to changing environments. In fact, meta-learning has been increasingly applied to the study of human behavior in recent years (Binz et al., 2023; Griffiths et al., 2019; Silvetti et al., 2022). Meta-learning models have been shown to capture a broad range of empirically observed behavior. For example, they can describe several properties of heuristic human decision-making (Binz et al., 2022) and have been shown to replicate human biases in probabilistic interpretation (Dasgupta et al., 2020). We are currently unaware of such datasets, but future studies should evaluate whether human participants can also flexibly adapt between adversarial and stable or volatile environments, and try and fit meta-learning models to assess if humans meta-learn some, if any, of the critical parameters in our model to achieve strategically variable behavior.

Here, we leveraged our model to study strategically variable behavior in humans, pigeons and rats in adversarial environments. The second-order entropy, characterizing the variability in responses, confirmed similar levels of variability generated by these subjects as generated by our model simulations in Study 1. All these species are capable of variable choice-making as a strategy, but the parameter estimates showed they achieve it in a different way. Our main model, the ε-α-λ model, showed the best fit.

The estimates showed high epsilon values for all species, and learning rates spanning almost the entire range of available values for humans and pigeons, but relatively high (>0.66) values for rats. The exact value for learning rate for rats was likely dependent on the precise reinforcement schedule, rather than species. Interestingly, we also observed a positive unchosen value-bias for humans, but negative ones for pigeons and rats. Taken together, this implies that the behavior of all species aligned with those of a stochastic generator. The behavior also seemed partially explained by that of a dynamic reinforcement and extinction process, especially for rats.

Finally, for humans, choices also seemed to rely on a frequency-based memory, where they strategically upregulated the values of actions that had not been performed for a while, something that pigeons and rats seemed to lack altogether. A first potential explanation for this is that humans could realize that variability in itself *is* a strategy, in a way (other) animals cannot. This awareness may be crucial for these memory processes, which ultimately may also lead to biases that deflect from randomness as observed in random number generation tasks (Ginsburg & Karpiuk, 1994; Joppich et al., 2004).

A second explanation could be that humans, as opposed to pigeons and rats, may be more likely to try and infer (a set of) rules from the environment that maximize their rewards (without necessarily *realizing* that variability in itself is the strategy). For example, Maes and colleagues (2015) studied generalization strategies in humans, pigeons and rats, and found that rat- and pigeon subjects generalized based on stimulus features, while most human participants showed generalization based on a set of inferred rules (Maes et al., 2015). In the adversarial environment of our study, it may be possible to infer such rules by keeping track of choice frequencies, as reflected in the positive λ-values. This type of decision-making process still leads to some degree of variability, be it biased and not completely random.

Finally, some non-human animals might even be unable to infer a rule based on memory. Indeed, (Lind & Jon-And, 2024) recently argued that non-human animals do not have memory for (stimulus) sequences (but see Inoue & Matsuzawa, 2007). Therefore, it is possible that they are less able to remember choice frequencies and use these to create more variable behavior. Instead, pigeons and rats might ‘give up’ easier when no actions are consistently rewarded and switch to a default setting that also results in spontaneous or unpredictable behavior (De Souza Barba, 2015). This does not require *realizing* that one should make variable choices, but could simply result from

a false belief that it does not matter what choice is made and a default stochastic strategy is most likely to help them learn new contingencies if they come up. This idea aligns with a putative stochastic generator mechanism and is reflected in the high epsilon values.

A clear avenue for future research is that our model currently models three potential cognitive mechanisms or learning dynamics, but more may be needed. The three mechanisms – a stochastic generator, dynamic reinforcement/extinction learning, and frequency-based memory – were inspired by prior literature on strategically variable behavior in environments where variability is selectively reinforced. We aimed to provide a first model to study and disentangle these different mechanisms within a single computational framework. However, it is possible, and not unlikely, that human and non-human animals may further rely on other mechanisms that could also be added to our model. For example, agents may use more complex rule-based strategies to generate variable behavior that currently went undetected in our model (e.g., Dayan & Niv, 2008; Dehaene et al., 2022; Lake et al., 2015), or selectively up- or down-regulate different learning rates for negative versus positive prediction errors (S. J. Gershman, 2015; Niv et al., 2012; Palminteri & Lebreton, 2022; Rosenbaum et al., 2022; Simoens et al., 2024; Wen et al., 2023).

Taken together, we provide a first computational framework to study strategically variable decision-making in adversarial environments, and show a clear distinction between the different underlying cognitive mechanisms that humans use to be strategically variable, and those employed by pigeons and rats. Our results are generally consistent with previous studies suggesting that humans show superior performance on problem-solving tasks that require self-control (MacLean et al., 2014). Importantly, the need for variable or unpredictable behavior in real-life is often confined to social situations, such as in games or sports, which humans may face more. In such

cases, it has been shown that humans often need to use more complex inferences that go beyond simpler forms of reinforcement learning (FeldmanHall & Nassar, 2021; Vélez & Gweon, 2021).

References

Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998). Random Generation and the Executive Control of Working Memory. *The Quarterly Journal of Experimental Psychology Section A*, *51*(4), 819–852. https://doi.org/10.1080/713755788

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Binz, M., Dasgupta, I., Jagadish, A., Botvinick, M., Wang, J. X., & Schulz, E. (2023). *Meta-Learned Models of Cognition*. http://arxiv.org/abs/2304.06729

Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022). Heuristics From Bounded Meta-Learned Inference. *Psychological Review*, *129*(5), 1042–1077. https://doi.org/10.1037/rev0000330

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Bryantt, D., & Church, R. M. (1974). The determinants of random choice*. In *Animal Learning & Behavior* (Vol. 2, Issue 4).

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review*, *67*(6), 380–400. https://doi.org/10.1037/H0040373

Capone, F., Capone, G., Ranieri, F., Di Pino, G., Oricchio, G., & Di Lazzaro, V. (2014). The effect of practice on random number generation task: A transcranial direct current stimulation study. *Neurobiology of Learning and Memory*, *114*, 51–57. https://doi.org/10.1016/j.nlm.2014.04.013

Dall, S. R. X., & Griffith, S. C. (2014). An empiricist guide to animal personality variation in ecology and evolution. *Frontiers in Ecology and Evolution*, *2*(FEB), 3. https://doi.org/10.3389/FEVO.2014.00003/BIBTEX

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A Theory of Learning to Infer. *Psychological Review*, *127*(3), 412–441. https://doi.org/10.1037/rev0000178

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. In *Current Opinion in Neurobiology* (Vol. 18, Issue 2, pp. 185–196). https://doi.org/10.1016/j.conb.2008.08.003

De Souza Barba, L. (2015). Controlling and Predicting Unpredictable Behavior. In *Behavior Analyst* (Vol. 38, Issue 1, pp. 93–107). Springer International Publishing. https://doi.org/10.1007/s40614-014-0019-9

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs: a hypothesis about human singularity. In *Trends in Cognitive Sciences* (Vol. 26, Issue 9, pp. 751–766). Elsevier Ltd. https://doi.org/10.1016/j.tics.2022.06.010

Denney, J., & Neuringer, A. (1998). Behavioral variability is controlled by discirminative stimuli. *Animal Learning & Behavior*, 154–162.

Donahoe, J. W., & Palmer, D. C. (1994). *Learning and complex behavior*. Allyn & Bacon.

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. In *Trends in Cognitive Sciences* (Vol. 25, Issue 12, pp. 1045–1057). Elsevier Ltd. https://doi.org/10.1016/j.tics.2021.09.002

Funamizu, A., Ito, M., Doya, K., Kanzaki, R., & Takahashi, H. (2012). Uncertainty in action-value estimation affects both action choice and learning rate of the choice behaviors of rats. *European Journal of Neuroscience*, *35*(7), 1180–1189. https://doi.org/10.1111/j.1460-9568.2012.08025.x

Gershman, S. (2021). *What Makes Us Smart*. Princeton University Press.

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*(5), 1320–1327. https://doi.org/10.3758/s13423-014-0790-3

Ginsburg, N., & Karpiuk, P. (1994). Random Generation: Analysis of the Responses. In *O Perceptual and Motor Skills* (Vol. 79).

Goris, J., Silvetti, M., Verguts, T., Wiersema, J. R., Brass, M., & Braem, S. (2021). Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *Autism*, *25*(2), 440–451. https://doi.org/10.1177/1362361320962237/ASSET/IMAGES/LARGE/10.1177_1362361320962237-FIG2.JPEG

Griffiths, T. L., Callaway, F., Chang, M. B., Grant, E., Krueger, P. M., & Lieder, F. (2019). Doing more with less: meta-reasoning and meta-learning in humans and machines. In *Current Opinion in Behavioral Sciences* (Vol. 29, pp. 24–30). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2019.01.005

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Bateson, M., Cools, R., Dukas, R., Giraldeau, L. A., Macy, M. W., Page, S. E., Shiffrin, R. M., Stephens, D. W., & Wolfe, J. W. (2015). Exploration versus exploitation in space, mind, and society. In *Trends in Cognitive Sciences* (Vol. 19, Issue 1, pp. 46–54). Elsevier Ltd. https://doi.org/10.1016/j.tics.2014.10.004

Humphries, D. A., & Driver, P. M. (1967). Erratic Display as a Device against Predators. *Science*, *156*(3783), 1767–1768. https://doi.org/10.1126/SCIENCE.156.3783.1767

Hutter, F., Kotthoff Lars, & Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges* (F. Hutter, L. Kotthoff, & J. Vanschoren, Eds.; 1st ed.). Springer Cham.

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. In *Current Biology* (Vol. 17, Issue 23). https://doi.org/10.1016/j.cub.2007.10.027

Jensen, G. (2018). Choice Under Conditions of Uncertainty: Methods for Analyzing the Structure of Behavior. *PsyArXiv Preprint*.

Jensen, G., Miller, C., & Neuringer, A. (2006). *Comparative cognition: Experimental Explorations of Animal Intelligence*.

Jin, F., Yang, L., Yang, L., Li, J., Li, M., & Shang, Z. (2024). Dynamics Learning Rate Bias in Pigeons: Insights from Reinforcement Learning and Neural Correlates. *Animals*, *14*(3), 489. https://doi.org/10.3390/ani14030489

Joppich, G., Däuper, J., Dengler, R., Johannes, S., Rodriguez-Fornells, A., & Münte, T. F. (2004). Brain potentials index executive functions during random number generation. *Neuroscience Research*, *49*(2), 157–164. https://doi.org/10.1016/j.neures.2004.02.003

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*(4–6), 549–559. https://doi.org/10.1016/S0893-6080(02)00048-5

Kilicay-Ergin, N. H., & Jablokow, K. W. (2012). Problem-solving variability in cognitive architectures. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *42*(6), 1231–1242. https://doi.org/10.1109/TSMCC.2012.2201469

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Laurent, V., & Balleine, B. W. (2015). Factual and Counterfactual Action-Outcome Mappings Control Choice between Goal-Directed Actions in Rats. *Current Biology*, *25*(8), 1074–1079. https://doi.org/10.1016/j.cub.2015.02.044

Lind, J., & Jon-And, A. (2024). A sequence bottleneck for animal intelligence and language? In *Trends in Cognitive Sciences*. Elsevier Ltd. https://doi.org/10.1016/j.tics.2024.10.009

Machado, A. (1989). Operant Conditioning of Behavioral Variability using a Percentile Reinforcement Schedule. *Journal of the Experimental Analysis of Behavior*, 155–166.

Machado, A. (1992). Behavioral Variability and Frequency-Dependent Selection. *Journal of the Experimental Analysis of Behavior*, 241–263.

Machado, A. (1993). Learning variable and stereotypical sequences of responses: Some data and a new model. In *Behavioural Processes*.

Machado, A., & Tonneau, F. (2012). Operant Variability: Procedures and Processes. *The Behavior Analyst*, 249–255.

MacLean, E. L., Hare, B., Nun, C. L., Addess, E., Amic, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N., Burkart, J. M., Call, J., Cantlo, J. F., Chek, L. G., … Zhao, Y. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20). https://doi.org/10.1073/pnas.1323533111

Maes, E., De Filippo, G., Inkster, A. B., Lea, S. E. G., De Houwer, J., D'Hooge, R., Beckers, T., & Wills, A. J. (2015). Feature- versus rule-based generalization in rats, pigeons and humans. *Animal Cognition*, *18*(6), 1267–1284. https://doi.org/10.1007/s10071-015-0895-8

Maye, A., Hsieh, C. H., Sugihara, G., & Brembs, B. (2007). Order in Spontaneous Behavior. *PLOS ONE*, *2*(5), e443. https://doi.org/10.1371/JOURNAL.PONE.0000443

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration-exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience*, *30*(37), 12366–12378. https://doi.org/10.1523/JNEUROSCI.0822-10.2010

Nergaard, S. K., & Holth, P. (2020). A Critical Review of the Support for Variability as an Operant Dimension. *Perspectives on Behavior Science*, *43*(3), 579–603. https://doi.org/10.1007/s40614-020-00262-y

Neuringer, A. (1986). Can People Behave "Randomly?": The Role of Feedback. In *Journal of Experimental Psychology: General* (Vol. 115, Issue 1).

Neuringer, A. (1991). Operant Variability and Repetition as Functions of Interresponse Time. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*(1), 3–12. https://doi.org/10.1037/0097-7403.17.1.3

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, *32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Oomens, W., Maes, J. H. R., Hasselman, F., & Egger, J. I. M. (2015). A time series approach to random number generation: Using recurrence quantification analysis to capture executive behavior. *Frontiers in Human Neuroscience*, *9*(JUNE). https://doi.org/10.3389/fnhum.2015.00319

Page, S., & Neuringer, A. (1985). Variability Is an Operant. In *Journal of Experimental Psychology: Animal Behavior Processes* (Vol. 11, Issue 3).

Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. In *Trends in Cognitive Sciences* (Vol. 26, Issue 7, pp. 607–621). Elsevier Ltd. https://doi.org/10.1016/j.tics.2022.04.005

Parsonson, B. S., & Baer, D. M. (1978). Training generalized improvisation of tools by preschool children1. *Journal of Applied Behavior Analysis*, *11*(3), 363–380. https://doi.org/10.1901/JABA.1978.11-363

Raza Ali, A., Budka, M., & Gabrys, B. (2020). *A Review of Meta-level Learning in the Context of Multi-component, Multi-level Evolving Prediction Systems*. https://doi.org/10.48550/arXiv.2007.10818

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement*. https://www.researchgate.net/publication/239030972

Rosenbaum, G. M., Grassie, H. L., & Hartley, C. A. (2022). Valence biases in reinforcement learning shift across adolescence and modulate subsequent memory. *ELife*, *11*. https://doi.org/10.7554/eLife.64620

Ross, B. M. (1955). Randomization of a Binary Series. *The American Journal of Psychology*, *68*(1), 136. https://doi.org/10.2307/1418397

Schweighofer, N., & Doya, K. (2003). Meta-learning in Reinforcement Learning. *Neural Networks*, *16*(1), 5–9. https://doi.org/10.1016/S0893-6080(02)00228-9

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 623–656.

Siegler, R. S. (1998). *Emerging Minds*. Oxford University Press.

Sikora, R. T. (2008). Meta-learning optimal parameter values in non-stationary environments. *Knowledge-Based Systems*, *21*(8), 800–806. https://doi.org/10.1016/j.knosys.2008.03.041

Silvetti, M., Lasaponara, S., Daddaoua, N., Horan, M., & Gottlieb, J. (2022). *A Reinforcement Meta-Learning Framework of Executive Function and Information Demand*.

Simoens, J., Verguts, T., & Braem, S. (2024). Learning environment-specific learning rates. *PLOS Computational Biology*, *20*(3), e1011978. https://doi.org/10.1371/journal.pcbi.1011978

Sutton, R. S., & Barto, A. G. (2020). *Reinforcement learning : an introduction*.

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*(1), 77–101. https://doi.org/10.1111/j.2044-8295.1998.tb02674.x

Uddin, L. Q. (2021). Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. In *Nature Reviews Neuroscience* (Vol. 22, Issue 3, pp. 167–179). Nature Research. https://doi.org/10.1038/s41583-021-00428-w

van Heeswijk, W. (2020). *A Minimal Working Example for Continuous Policy Gradients in TensorFlow 2.0*. https://towardsdatascience.com/a-minimal-working-example-for-continuous-policy-gradients-in-tensorflow-2-0-d3413ec38c6b

Vélez, N., & Gweon, H. (2021). Learning from other minds: an optimistic critique of reinforcement learning models of social learning. In *Current Opinion in Behavioral Sciences* (Vol. 38, pp. 110–115). Elsevier Ltd. https://doi.org/10.1016/j.cobeha.2021.01.006

Verbeke, P., & Verguts, T. (2024). Humans adaptively select different computational strategies in different learning environments. *Psychological Review*. https://doi.org/10.1037/rev0000474

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. In *Psychological Bulletin* (Vol. 77, Issue 1).

Walker, M., & Wooders, J. (2001). *Minimax Play at Wimbledon*.

Wen, T., Geddert, R. M., Madlon-Kay, S., & Egner, T. (2023). Transfer of Learned Cognitive Flexibility to Novel Stimuli and Task Sets. *Psychological Science*, *34*(4), 435–454. https://doi.org/10.1177/09567976221141854

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3–4), 229–256. https://doi.org/10.1007/BF00992696

Wilson, R. C., & Collins, A. G. (2019). *Ten simple rules for the computational modeling of behavioral data*.