

# Flexibility of learning in complex worlds

Olof Leimar<sup>1, \*</sup>, Andrés E Quiñones<sup>2</sup>, and Redouan Bshary<sup>3</sup>

<sup>1</sup>Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden

<sup>2</sup>Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia

<sup>3</sup>Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland

\*Corresponding author; e-mail: olof.leimar@zoologi.su.se



## Abstract

Learning to adjust to changing environments is an important aspect of behavioural flexibility. Here we investigate the possible advantages of flexible learning rates in volatile environments, using learning simulations. We compare two established learning mechanisms, one with fixed learning rates (Rescorla-Wagner) and one with flexible rates (Autostep). We study three types of volatility: transitions from a simpler to a more complex foraging environment, reversal learning, and learning set formation. For transitions to a complex world, we use developing cleaner fish as an example, having more types of client fish to choose between as they become adult. There are other similar transitions in nature, such as migrating to a new and different habitat. Performance in reversal learning and in learning set formation are commonly used experimental measures of behavioural flexibility. Concerning transitions to a complex world, we show that both Rescorla-Wagner and Autostep learn to make rewarding choices, losing only 5-10% of maximum available rewards in the period after a transition, but Autostep performs noticeably better than Rescorla-Wagner. For reversal learning, we show that Autostep sharply increases learning rates with each successive reversal, thus increasing its performance, whereas there is no such effect for Rescorla-Wagner. For learning set formation, we find no effect of increasing performance with successive shifts to new stimuli to discriminate for either flexible or fixed learning rates. Flexible learning rates might thus explain increasing performance in reversal learning, but not in learning set formation. We discuss our results in relation to current ideas about behavioural flexibility.

*Keywords:* Rescorla-Wagner learning, Autostep, meta learning, volatility, stochasticity, prediction error, reversal learning, learning set formation.

# 1 Introduction

2 The ability of animals to adjust to new and complex environments through learning  
 3 is an important aspect of adaptive behavioural flexibility. In animal psychology and  
 4 behavioural ecology, different meanings have been given to the term behavioural flexi-  
 5 bility (Audet and Lefebvre 2017; Lea et al. 2020), but here we are concerned with the  
 6 ability to adjust to environmental change using learning. The question we ask is how  
 7 well different mechanisms of reinforcement learning, with either constant or variable  
 8 learning rates, serve to adapt behaviour in a volatile environment. Specifically, we  
 9 investigate how big the advantage of having flexible learning rates might be when the  
 10 environment is volatile.

11 It is known from neuroscience studies that humans and other animals adjust learning  
 12 rates to the volatility of rewards (e.g., Behrens et al. 2007; Diederer and Schultz 2015;  
 13 Grossman et al. 2022). Performance in reversal learning is one measure that has been  
 14 used to describe behavioural flexibility (Deaner et al. 2006; Bond et al. 2007; Izquierdo  
 15 et al. 2017; Liu et al. 2016; Buechel et al. 2018; Boussard et al. 2021; Triki et al. 2022),  
 16 and this performance might be improved by flexible learning rates. The performance in  
 17 other measures of behavioural flexibility, such as learning set formation (set-shifting),  
 18 where an individual encounters a sequence of novel discrimination tasks (Harlow 1949;  
 19 Wilson et al. 1985; Bailey et al. 2007), could conceivably also be enhanced by flexible  
 20 learning rates.

21 To investigate the potential advantages of variable learning rates, we use the learn-  
 22 ing mechanism by Rescorla and Wagner (1972) as a baseline. Rescorla-Wagner learning  
 23 assumes constant learning rates for each stimulus dimension and is among the most in-  
 24 vestigated approaches to learning. Learning rates can differ between dimensions, but for  
 25 a given dimension the rate is fixed. While being a strong candidate for adaptive learn-  
 26 ing, it is known that Rescorla-Wagner is not strictly optimal in volatile environments  
 27 (Dayan et al. 2000; Trimmer et al. 2012). Several alternatives to Rescorla-Wagner  
 28 learning have been proposed, typically involving flexible learning rates. In our compar-  
 29 isons here, we use a learning algorithm called Autostep (Mahmood et al. 2012), because  
 30 of its robustness in adapting learning rates without the need for extensive tuning of  
 31 parameters. It is a further refinement of the so-called delta-bar-delta algorithm (Jacobs  
 32 1988; Sutton 1992a, 2022), and it falls into the category of meta-learning approaches  
 33 (Sutton 2022), i.e. learning to learn.

34 In the following, we outline the learning models and simulations we use, and then  
 35 present results from different situations where flexible learning might be advantageous.  
 36 The first cases we investigate are inspired by the situation for developing cleaner fish,

as they become **adult** and transition from a simpler to a more complex set of client fish species to choose between and clean (Triki et al. 2019). More generally, these cases illustrate challenges encountered by many learning animals. Examples include migrants experiencing a shift to a new and different foraging environment (Bairlein and Simons 1995; Pierce and McWilliams 2005), and seasonal changes that expose a forager to new food types (Janmaat et al. 2016). We also compare the performance of Rescorla-Wagner learning to Autostep for a case of reversal learning, extended over several reversals, and we make a similar comparison for a case of learning set formation. Finally, we give an illustration of the performance of Rescorla-Wagner and Autostep when the degree of stochasticity of the environment becomes high. Although the impact of the relative strengths of stochasticity and volatility has not been emphasized in experimental psychology or behavioural ecology, it has been dealt with in neuroscience (Nassar et al. 2010; Piray and Daw 2021). The general conclusion is that stochasticity should favour lower learning rates, allowing a learner to average over more trials.

We discuss our results in relation to existing ideas about the significance of flexible learning, making the point that relatively simple mechanisms of adjustment of learning rates could, wholly or partially, explain some of the observed phenomena of flexibility of learning. Such adjustments could represent specific adaptations to environmental volatility, or they could be consequences of broader cognitive adaptations, for instance relating to attention and memory. We also comment on the Autostep approach in relation to previous and current ideas in the field (e.g., Mackintosh 1975; Pearce and Hall 1980; Pearce and Mackintosh 2010; Holland and Schiffino 2016; Soltani and Izquierdo 2019). Finally, we argue that performance in the face of environmental volatility, including transitions to new and complex environments, is a good candidate for the selective advantage of flexible learning.

## Learning models and approaches

The kind of learning we study is where an individual learns the values of stimuli that can be distinguished by certain characteristics, which we refer to as features of compound stimuli. The characteristics define different stimulus dimensions, which can be things like the colour, texture, or shape of potential food items encountered by a forager. For the cleaner fish example, a compound stimulus would be a client fish. **There are different ways that individuals might estimate the values of compound stimuli, one being to form an entirely separate estimate for each type of stimulus (sometimes called ‘object learning’, Farashahi et al. (2020)).**

71 We focus on another mechanism, in which an individual learns an estimate of a  
 72 value for each feature of a compound stimulus, and uses the sum of these values to  
 73 estimate the value of the compound. The approach corresponds to long-standing ideas  
 74 about classical conditioning in experimental psychology, when animals respond to the  
 75 component stimuli that are present in a learning trial. The case most frequently studied  
 76 is that of absence/presence features (0/1 stimulus components), where a feature has  
 77 only two states, either being absent or present in the compound, and we make use of this  
 78 in our learning simulations. There are of course other cases, for instance quantitative  
 79 stimulus dimensions, and we include overall stimulus size as one such dimension.

80 Perhaps the most influential formulation of these ideas is the learning mechanism  
 81 proposed by Rescorla and Wagner (1972). In their approach, if  $w_m$  is an individual's  
 82 current estimate of the value of a certain absence/presence feature from stimulus di-  
 83 mension  $m$ , and the feature is present in a learning trial, the individual updates its  
 84 estimate to  $w'_m$ , where

$$w'_m = w_m + \alpha_m(R - Q). \quad (1)$$

85 Here  $R$  is the reward perceived by the individual from interacting with the compound,  
 86 and  $Q$  is the individual's previous estimate of the value of the compound. The quantity

$$\delta = R - Q \quad (2)$$

87 is referred to as the prediction error, and is the difference between the reward  $R$  cur-  
 88 rently experienced by the individual and its prior estimate  $Q$  of the reward. The change  
 89 in the estimated feature value  $w_m$  in equation (1) is thus the learning rate  $\alpha_m$  times  
 90 the prediction error, and tends to move the estimate towards the true value. Learning  
 91 rates could differ between stimulus dimensions and could also change over time. The  
 92 main question we ask is how big the advantage of flexible learning rates might be.

93 If  $x_m$  indicates the feature from stimulus dimension  $m$ , so that  $x_m$  is a 0/1 variable,  
 94 the estimate of the value of the compound is given by the sum of all feature values that  
 95 are present:

$$Q = \sum_{m=1}^M x_m w_m. \quad (3)$$

96 This formula also applies to quantitative stimulus dimensions, for which  $w_m$  is the  
 97 estimated reward per unit of the dimension. For simplicity, we limit ourselves to  
 98 additive reward structures, although there are other cases that occur in nature, such

as when features interact in indicating the value. There is also random variation in rewards. For instance, for client fish visiting cleaners, there is work showing that the number of parasites that cleaners remove and feed on is correlated with client size, but the correlations are not extremely high (Grutter 1994, 1995).

When an individual can choose between two compound stimuli with estimated values of  $Q_1$  and  $Q_2$ , we assume that the individual chooses stimulus 1 with probability

$$p_1 = \frac{1}{1 + \exp(-\omega(Q_1 - Q_2))}, \quad (4)$$

where  $\omega$  is a parameter (we used  $\omega = 5$  in our simulations; see the curve in Figure 1b below). This is referred to as a soft-max rule, going from estimated values to a choice, and it is commonly used in reinforcement learning models (e.g., Sutton and Barto 2018).

Rescorla and Wagner (1972) assumed that learning rates stay constant over time, but there are a number of suggestions for how they might vary. One idea is that an increase in prediction errors could indicate to an individual that it should change its learning rates. We will investigate how much better an individual with flexible learning rates is at selecting higher value compounds, compared to a Rescorla-Wagner learner.

The Autostep method (Mahmood et al. 2012) is a meta-learning approach (Sutton 2022) that adjusts learning rates based on the recent prediction-error history. An overall idea of such meta-learning algorithms is to adjust learning rates in a way that minimizes prediction errors. Autostep is a further refinement of the delta-bar-delta method (Sutton 1992a), making it more robust. The intuition behind delta-bar-delta is to increase a learning rate  $\alpha_m$  if in recent trials the estimate  $w_m$  has been increasing (decreasing) and the current prediction error indicates an additional increase (decrease) in  $w_m$ , corresponding to a positive correlation between the recent and current changes to  $w_m$ . Similarly, the learning rate is decreased for a negative correlation, because this indicates that the current change in  $w_m$  overshoots the true value. The algorithm also changes learning rates on the log-scale, which allows for a fairly large range of values for the rates. These properties of IDBD also hold for Autostep. More details on the learning algorithms we use appears in the supplements.

## Learning simulations

As mentioned, the first learning environments we use in our simulations are inspired by the situation for developing cleaner fish, as they become adult and transition from a simpler to a more complex set of client fish species to choose between and clean. Here

we describe the simulations for such cases, where there is a transition from a simpler to a more complex learning environment.

### Stimulus dimensions and compound stimuli

In order to characterise many (up to 10) different compound stimuli, there are 10 stimulus dimensions. The first four dimensions are as follows, together with their true values.

1. The first dimension,  $x_1$ , is quantitative, like client size, and has a positive true value,  $W_1 = 1.0$ .
2. The second dimension is 0/1, and has a zero true value,  $W_2 = 0$ , so it is an irrelevant dimension.
3. The third dimension is 0/1 and has a positive true value,  $W_3 = 1.0$ , so it is a relevant dimension.
4. The fourth dimension is 0/1 and has a negative true value,  $W_4 = -1.0$ , so it is also a relevant dimension.

An additional six 0/1 dimensions are described in Table 1 below. From combinations of the four first dimension we have four types of compound stimuli (e.g., corresponding to four client species).


1. The first type has small size,  $x_1 = \exp(y_1 + z_x)$ , with  $z_x$  normally distributed with mean zero and standard deviation  $\sigma_x = 0.25$ , and  $y_1$  so that  $\bar{x}_1 = \bar{x}_{\text{small}} = 1$  (which happens for  $y_1 = -\sigma_x^2/2$ ), and absence of features in the other dimensions. This could be a species of small clients.
2. The second type has large size,  $x_1 = \exp(y_2 + z_x)$ , again with  $z_x$  normally distributed with mean zero and standard deviation  $\sigma_x$ , and  $y_2$  such that  $\bar{x}_1 = \bar{x}_{\text{large}} = 2$ , and presence of a feature in the second dimension, and absence of features in the other dimensions. This could be a species of large clients that are characterised by a feature  $x_2$  that is irrelevant for reward (size is sufficient to predict reward).
3. The third type of compound stimulus is the same as the second for the first two dimensions, but it has a feature present in the third dimension and no feature in the fourth. This is then a species of more valuable large clients. Perhaps parrotfish could be an example

162 4. The fourth type of compound stimulus is the same as the second for the first two  
 163 dimensions, and it has no feature in the third dimension but a feature present in  
 164 the fourth. This is then a species of less valuable large clients. Perhaps damselfish  
 165 could be an example.

166 These compound stimuli, together with six additional compound stimuli, are described  
 167 in Table 1. Note that we assume log-normal distributions for the first stimulus di-  
 168 mension and also for the stochasticity of rewards, in order to ensure that values are  
 169 positive.

## 170 Learning trials

171 We first consider two cases of sequences of learning trials. In both cases, there is an  
 172 initial phase of  $T$  trials of learning ( $T = 1000$ ) with only the first two compound stimuli  
 173 (e.g., one species of small clients and one species of large clients). This is followed by  
 174 a phase of an additional  $T$  trials of learning in a more complex world. In case 1,  
 175 individuals learn to discriminate between the first four compound stimuli in Table 1  
 176 (which could be four client species). In case 2, the world is even more complex, such  
 177 that individuals learn to discriminate all 10 compound stimuli in Table 1. In both  
 178 cases, an individual can choose between two compound stimuli in each trial, and these  
 179 are randomly drawn from all types that occur in that phase of learning of that case.

180  For reversal learning, there is first a phase of 100 trials where individuals can choose  
 181 between a rewarded stimulus ( $R = 1$ ), with a feature present in dimension 1, and an  
 182 unrewarded stimulus ( $R = 0$ ), with no feature in dimension 1 but a feature present in  
 183 dimension 2. In practice, the discrimination could be between blue and green stimuli.  
 184 These 100 trials are enough for individuals to learn to prefer the rewarded stimulus.  
 185 In the next 100 trials the rewards are reversed. The entire procedure is then repeated  
 186 for another 200 trials, i.e. an additional two reversals.

187 For learning set formation, the first 100 trials are as in the reversal learning case,  
 188 but in subsequent intervals of 100 trials, entirely new pairs of rewarded and unrewarded  
 189 stimuli are used, with features in new stimulus dimensions. A total of four pairs are  
 190 used, making up a total of 400 trials. As an example, the four pairs could be blue and  
 191 green stimuli, followed by circular and square stimuli, followed by striped and plain  
 192 stimuli, followed by horizontally and vertically oriented stimuli.

193 For these cases, we present results based on replicate simulations of learning for  
 194 100 individuals. We assume that the reward from a compound stimulus has a log-  
 195 normal distribution around the true expected value, with a standard deviation  $\sigma_R$  on  
 196 the log scale. For the transitions to a more complex world, we use  $\sigma_R = 0.10$  (in the



197 supplements, we show results for higher stochasticity,  $\sigma_R = 0.50$ ; cases 3 to 6), and for  
 198 reversal learning and learning set formation, we use  $\sigma_R = 0.02$ .

199 As the starting value of learning rates, we use  $\alpha_m = 0.04$ , which allows for learning  
 200 of unit value differences over 50 to 100 trials. For the starting estimated values, we  
 201 used  $w_m = 0$ ; this might hold for individuals without any previous experience of the  
 202 stimulus dimension.

## 203 Results

204 The first phase of learning for cases 1 and 2, with only two types of compound stimuli, is  
 205 illustrated in Figure 1. The variation in rewards, shown in Figure 1a, comes both from  
 206 random variation in the first stimulus dimension (e.g., client size), and from random  
 207 variation in rewards from a client with a given true expected reward. The sigmoid  
 208 soft-max curve from equation (4) appears in panel b, and the learning rates  $\alpha_m$  and  
 209 estimated values  $w_m$  are shown in panels c and d, averaged over 100 replicates and  
 210 blocks of 10 learning trials. As seen in Figure 1d, Rescorla-Wagner and Autostep have  
 211 similar performance in the first phase of learning, with only a slight advantage for  
 212 Autostep in achieving better estimates of the true values.

213 There are two cases for the second phase of learning, and the outcome of these  
 214 learning simulations is illustrated in Figure 2. The learning rates for Autostep increase  
 215 sharply in the second phase (Figure 2a), especially for case 2 where many new stimulus  
 216 dimensions are needed for discrimination, whereas the rates for Rescorla-Wagner stay  
 217 constant. Another comparison of the performance of Rescorla-Wagner and Autostep  
 218 for the two simulated cases appears in Figure 3. In this figure the performance is  
 219 measured in terms of the deviation of an individual’s estimate from the true value,  
 220 implemented as the root mean square error (RMSE). Autostep does noticeably better  
 221 than Rescorla-Wagner in reducing the errors in the value estimates but, as seen from  
 222 the Figures 1 and 2, RMSE is not the only thing that matters. Thus, even if a learner  
 223 deviates in its estimates, it can still be the case that it makes a correct choice between  
 224 two compound stimuli, because the deviations might be similar for the two stimuli.  
 225 Quantitatively, over the first 250 post-change trials for case 2, Autostep has a loss of  
 226 7.5% of maximum possible reward per trial, whereas Rescorla-Wagner has a higher  
 227 loss of 11.0%. Over the first 500 post-change trials, these losses are 4.1% and 6.6%.  
 228 Thus, Autostep does better than Rescorla-Wagner in handling a transition to a more  
 229 complex world, but the differences are moderate, and not dramatic, seen over timescales  
 230 of several hundreds of trials.



231 A different type of analysis of transitions to a more complex world is to consider how  
 232 much an individual who fails **learn** anything new about the more complex world would  
 233 lose in terms of rewards. For our cases 1 and 2, this would mean that individuals base  
 234 their choices only on compound stimulus size, also after the transition. Quantitatively,  
 235 for case 1, where the new world is only moderately more complex, using only size to  
 236 choose in the second phase would result in a reward loss of around 7.4% per trial, and  
 237 the corresponding figure for case 2 is around 22% per trial, which is a substantial loss.  
 238 Note that these losses would apply to all 1000 trials in the second phase, and would be  
 239 approximately the same for (appropriately modified versions of) Rescorla-Wagner and  
 240 Autostep. It follows that fairly large advantages can be gained by learning about the  
 241 new stimulus dimensions in the more complex world.

242 A comparison of the performance of Rescorla-Wagner and Autostep in reversal  
 243 learning appears in Fig. 4. For Autostep, the learning rates increase sharply with each  
 244 successive change in rewards (Figure 4a), whereas for Rescorla-Wagner the learning  
 245 rates are constant. A consequence of this is that Autostep increases its performance  
 246 over successive reversals (Figure 4b, c). This means that learning-rate flexibility, for  
 247 instance as implemented by Autostep, could contribute to observed increases in per-  
 248 formance over successive episodes of reversal learning.

249 Figure 5 shows a similar comparison of the performance in learning set formation. In  
 250 this case there is no additional increase in learning rates for Autostep over successive  
 251 shifts in pairs of stimuli (Figure 5a), and consequently no increase in performance  
 252 (Figure 5b, c). Thus, in contrast to reversal learning, learning-rate flexibility does not  
 253 increase the performance in learning set formation over successive shifts in stimuli to  
 254 discriminate.

255 Finally, in the supplements we analyse cases similar to those in Figures 1-3, but with  
 256 high stochasticity (Figures S1-S3). Compared to the cases with lower stochasticity,  
 257 the learning rates for Autostep are lower, resulting in better estimates of the true  
 258 values (Figure S3), but there are no dramatic additional advantages for Autostep over  
 259 Rescorla-Wagner in gaining rewards (Figure S2). Extending the phases of learning to  
 260  $T = 10000$  trials (Figures S4-S6), the lowering of learning rates for Autostep is even  
 261 more pronounced, in particular in the first phase of learning (Figure S4). As a result,  
 262 Autostep does much better than Rescorla-Wagner in estimating the true values (Figure  
 263 S6).

## Discussion

In our comparisons of flexible (Autostep) and fixed (Rescorla-Wagner) learning rates, we found pronounced variation in learning rates for the Autostep algorithm (Figures 1c, 2a, 4a, 5a). As a consequence, Autostep performed better than Rescorla-Wagner in estimating the true values of different stimulus dimensions (Figures 1d, 2b, 3, 4b, 5b). For our simulated cases of transitions from a simpler to a more complex world, this meant that Autostep made more correct choices (Figure 2c) and achieved higher post-transition rewards (Figure 2d) than was the case for Rescorla-Wagner. The effects of variable rates on rewards were moderate, but might still be large enough for this kind of learning flexibility to evolve. Alternatively, flexible learning rates could be an aspect of broader cognitive adaptations, relating to attention, memory, and the handling of environmental and social complexity (Emery and Clayton 2004; Deaner et al. 2006; Bond et al. 2007; Izquierdo et al. 2017; Rmus et al. 2021; Leimar et al. 2022).

Our analysis of two frequently used measures of behavioural flexibility produced contrasting results. We found that Autostep increased its performance in reversal learning with each switch in rewards (Figure 4c). This means that a mechanism similar to the one causing the learning rates of Autostep to increase – involving sensitivity to prediction errors that consistently change estimated values over several trials – might contribute to observed improvements in performance in reversal learning with successive switches. In contrast, we found no similar increase in the performance of Autostep over successive shifts of stimuli in learning set formation (Figure 5c). The reason is that, when there is an entirely new situation, in the sense that all stimulus components that a learner encounters come from new stimulus dimensions, the learning rates for Autostep for these dimensions start from scratch. The same ought to hold for other learning mechanisms that do not increase learning rates for stimulus dimensions that a learner so far has not encountered. A tentative conclusion is that increasing performance in reversal learning and in learning set formation correspond to distinct cognitive capacities.

The possibility of different learning rates for different stimulus dimensions is an important aspect of the Rescorla-Wagner model. Its original aim was to explain phenomena such as overshadowing and blocking (Rescorla and Wagner 1972; Miller et al. 1995), and overshadowing of one stimulus component by another depends on differences in learning rates. This is often described in terms of the salience or associability of stimulus components. In nature, the perceived salience of different stimulus components might be adaptative for a particular group of animals. For instance, for some birds the colour of artificial prey is more salient than the shape (Kazemi et al. 2014),

and such higher learning rates for colour might be adaptive. It is learning rate constancy over time, not over stimulus dimensions, that holds for the Rescorla-Wagner model. Our assumption of the same Rescorla-Wagner learning rate for different stimulus dimensions is thus not at all necessary, but is used as a convenient default in the comparison with Autostep.

## Learning models

Many learning models have been proposed in the literature, apart from the ones we study here. Some were developed by experimental psychologists and focus on classical conditioning (e.g., Mackintosh 1975; Pearce and Hall 1980; Le Pelley 2010; Pearce and Mackintosh 2010; Esber and Haselgrove 2011). Although these approaches contain interesting and influential ideas, they turn out not to be suitable for our learning simulations here. The reason is that the specific algorithms have difficulties handling large numbers of stimulus dimension and, furthermore, only allow for fairly limited variation in learning rates.

These approaches discuss variation in learning rates in terms of effects of attention on learning. The idea that attention to stimulus components could be important for learning is often put forward and has been investigated experimentally (Beesley et al. 2015; Niv et al. 2015; Leong et al. 2017; Tomkowiak-Rodas et al. 2021). Nevertheless, models with variation learning rates need not explicitly include attention as a mechanism (Dayan et al. 2000), and Autostep is an example of this.

There are also Kalman-filter-inspired learning models (the Kalman filter originated in the engineering related fields of optimisation and control). Examples are described by Sutton (1992b), Dayan et al. (2000), and Gershman (2015). The Kalman filter gives an optimal solution to a control problem, in certain mathematically well defined situations. It can be used to construct optimal learning algorithms in certain cases where the relative magnitudes of volatility and stochasticity are known (Dayan et al. 2000; Gershman 2015; Piray and Daw 2021). In many situations where the Kalman filter is optimal, the IDBD algorithm achieves approximately the same performance (Sutton 1992b). Because the Autostep algorithm is a more robust version of IDBD, it is reasonable to expect that it has approximately the same performance as a Kalman filter model in situations where the Kalman filter is optimal. A seeming advantage for algorithms such as Autostep and IDBD over a Kalman filter is that they do not require a priori knowledge of the relative magnitudes of volatility and stochasticity.

There is much work in theoretical neuroscience on neural-network-based learning models. This work is of interest if it helps in identifying neural correlates of learning

phenomena. An influential example is the modelling by Wang et al. (2018). They present a general perspective on meta learning and report on learning simulations for situations similar to reversal learning and learning set formation. In one simulation they trained a network to obtain rewards in situations with changing volatility, and the network then showed higher learning rates for higher reward volatility, in a similar way as was found in an experiment by Behrens et al. (2007). In another simulation a network was trained on learning set formation, and subsequently showed increasing performance similar to what was found in the original experiments by Harlow (1949). These are interesting results, but it is not clear which kinds of cognitive mechanisms caused the networks to succeed in the learning tasks.

### Behavioural flexibility

The idea that behavioural flexibility should be adaptive in complex worlds is well established. There is evidence that animals that are known or believed to have the cognitive capacities associated with a larger brain, and thus presumably show more flexible behaviour, are more successful in novel environments (Sol et al. 2002, 2005). Conversely, there is evidence suggesting that invasive species have cognitive abilities that allow flexible behaviour (Szabo et al. 2020). Among the many examples of ecologically relevant situations where there can be a shift to a more complex world are invasions into new habitats, but also new contextual cues for food choice (Hansen et al. 2010);

For learning-rate flexibility, which is the focus of our investigation here, it is worth noting that flexibility does not only entail higher learning rates for higher reward volatility, but also lower learning rates for higher reward stochasticity (as seen by comparing Figures 1 and 2 with S1, S2, S4, and S5). While effects of reward stochasticity have been investigated in neuroscience (Nassar et al. 2010; Piray and Daw 2021), there seems to be a lack of studies by behavioural ecologists.

Reversal learning experiments, in particular those involving serial reversals, can detect increasing learning rates with increasing reward volatility, as illustrated by our Figure 4. One focus of such studies has been whether species or groups of species differ in this performance. For instance, Bond et al. (2007) compared three corvid species, each of which showed increasing performance with successive reversals, but to different degrees, and suggested that differences in social complexity could explain the observation. Another example is that, based on overviews of several studies, there appears to be a pattern of little or no increase in performance over successive reversals in species of fish (Bitterman 1975; Boussard et al. 2020), in contrast to what is found

in other groups of vertebrates. There is so far no well established explanation for this possible difference. In our learning simulations of transitions to a more complex world, we used cleaner fish as an illustrative example, but up to now there are no serial reversal learning experiments on these, and it is not known to what extent they show flexible learning rates. Thus, in principle, cleaner fish learning might be better described by the Rescorla-Wagner model, for instance with adaptive but not very flexible learning rates for different stimulus dimensions, than by the Autostep model. Additional experiments are needed to settle the issue.

A number of studies have examined the neural correlates of learning rate flexibility. One general conclusion is that for humans and non-human primates, as well as for rodents, regions in the prefrontal cortex are important for reversal learning (Izquierdo et al. 2017), with serotonin neurons playing a role (Grossman et al. 2022). In fish, learning experiments on selection lines have shown that brain size influences performance in reversal learning (Buechel et al. 2018) and its decrease with age (Boussard et al. 2021), and that specifically relative telencephalon size influences the performance in reversal learning (Triki et al. 2022).

Our simulations showed a qualitative difference between performance in serial reversal learning (Figure 4) and in learning set formation (Figure 5), consistent with the idea that these depend on distinct cognitive capabilities. The capacity for learning set formation appears to have a considerably more narrow phylogenetic distribution, being largely restricted to primates (Harlow 1949; Warren 1966; Deaner et al. 2006) and some species of birds (Wilson et al. 1985; Emery and Clayton 2004; de Mendonça-Furtado and Ottoni 2008), than has increasing performance in serial reversal learning. Learning set formation is sometimes described as rule learning, with the rule being ‘win-stay, lose-shift’ (Warren 1966; Mackintosh et al. 1968; Emery and Clayton 2004), but the actual cognitive mechanism involved is not known. Furthermore, seemingly abstract rules that an experimenter has imposed (e.g., ‘precisely one out of two possibilities is rewarded’) need not correspond to important situations encountered in nature. It might be more important for animals to learn rules about categories of compound stimuli, for instance ‘predator’ and ‘non-predator’. Cleaner fish appear to use this categorisation to solve a problem of ‘avoiding punishment’ (Wismer et al. 2016).

Overall, our simulations show that adaptive learning rate flexibility can rely on relatively simple mechanisms, such as using correlations between current and recent changes in estimated values to adjust rates, as for Autostep. To the extent that our examples of transitions to a complex world are biologically realistic, one can also conclude that learning rate flexibility gives a clear but only moderately large advantage over fixed rates. In comparison, as we have shown, it is considerably more important to

learn at all about the new and informative stimulus dimensions in the complex world. Cognitive capacities allowing individuals to achieve this seem essential for behavioural flexibility, and might involve attention, memory, and exploration, in addition to flexible learning rates.

## Funding

This work was supported by a grant (2018-03772) from the Swedish Research Council to OL and a grant (310030\_192673/1) from the Swiss National Science Foundation to RB.

## Acknowledgements

We thank Annika Boussard for helpful comments.

## Conflict of interest

The authors declare no conflict of interest.

## Code availability

C++ source code for the individual-based simulations is available at GitHub, together with instructions for compilation on a Linux operating system:

## References

- Audet, J.-N. and Lefebvre, L. 2017. What’s flexible in behavioral flexibility? *Behavioral Ecology*, 28(4):943–947.
- Bailey, A. M., McDaniel, W. F., and Thomas, R. K. 2007. Approaches to the study of higher cognitive functions related to creativity in nonhuman animals. *Methods*, 42(1):3–11.
- Bairlein, F. and Simons, D. 1995. Nutritional adaptations in migrating birds. *Israel Journal of Ecology and Evolution*, 41(3):357–367.
- Beesley, T., Nguyen, K. P., Pearson, D., and Le Pelley, M. E. 2015. Uncertainty and predictiveness determine attention to cues during human associative learning. *Quarterly Journal of Experimental Psychology*, 68(11):2175–2199.

- 433 Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. 2007.  
 434 Learning the value of information in an uncertain world. *Nature Neuroscience*,  
 435 10(9):1214–1221.
- 436 Bitterman, M. E. 1975. The comparative analysis of learning. *Science*, 188(4189):699–  
 437 709.
- 438 Bond, A. B., Kamil, A. C., and Balda, R. P. 2007. Serial reversal learning and the  
 439 evolution of behavioral flexibility in three species of North American corvids ( *Gym-*  
 440 *norhinus cyanocephalus*, *Nucifraga columbiana*, *Aphelocoma californica*). *Journal of*  
 441 *Comparative Psychology*, 121(4):372–379.
- 442 Boussard, A., Amcoff, M., Buechel, S. D., Kotrschal, A., and Kolm, N. 2021. The  
 443 link between relative brain size and cognitive ageing in female guppies (*Poecilia*  
 444 *reticulata*) artificially selected for variation in brain size. *Experimental Gerontology*,  
 445 146:111218.
- 446 Boussard, A., Buechel, S. D., Amcoff, M., Kotrschal, A., and Kolm, N. 2020. Brain  
 447 size does not predict learning strategies in a serial reversal learning test. *Journal of*  
 448 *Experimental Biology*, 223(15):jeb224741.
- 449 Buechel, S. D., Boussard, A., Kotrschal, A., van der Bijl, W., and Kolm, N. 2018.  
 450 Brain size affects performance in a reversal-learning test. *Proceedings of the Royal*  
 451 *Society B: Biological Sciences*, 285(1871):20172031.
- 452 Dayan, P., Kakade, S., and Montague, P. R. 2000. Learning and selective attention.  
 453 *Nature Neuroscience*, 3(11):1218–1223.
- 454 de Mendonça-Furtado, O. and Ottoni, E. B. 2008. Learning generalization in problem  
 455 solving by a blue-fronted parrot (*Amazona aestiva*). *Animal Cognition*, 11(4):719–  
 456 725.
- 457 Deaner, R. O., van Schaik, C. P., and Johnson, V. 2006. Do some taxa have bet-  
 458 ter domain-general cognition than others? A meta-analysis of nonhuman primate  
 459 studies. *Evolutionary Psychology*, 4(1):147470490600400114.
- 460 Diederer, K. M. J. and Schultz, W. 2015. Scaling prediction errors to reward variability  
 461 benefits error-driven learning in humans. *Journal of Neurophysiology*, 114(3):1628–  
 462 1640.
- 463 Emery, N. J. and Clayton, N. S. 2004. The mentality of crows: convergent evolution  
 464 of intelligence in corvids and apes. *Science*, 306(5703):1903–1907.



- Esber, G. R. and Haselgrove, M. 2011. Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, 278(1718):2553–2561.
- Farashahi, S., Xu, J., Wu, S.-W., and Soltani, A. 2020. Learning arbitrary stimulus-reward associations for naturalistic stimuli involves transition from learning about features to learning about objects. *Cognition*, 205:104425.
- Gershman, S. J. 2015. A unifying probabilistic view of associative learning. *PLOS Computational Biology*, 11(11):e1004567.
- Grossman, C. D., Bari, B. A., and Cohen, J. Y. 2022. Serotonin neurons modulate learning rate through uncertainty. *Current Biology*, 32(3):586–599.e7.
- Grutter, A. S. 1994. Spatial and temporal variations of the ectoparasites of seven reef fish species from lizard island and heron island, australia. *Marine Ecology Progress Series*, 115:21–30.
- Grutter, A. S. 1995. Relationship between cleaning rates and ectoparasite loads in coral reef fishes. *Marine Ecology Progress Series*, 118:51–58.
- Hansen, B. T., Holen, O. H., and Mappes, J. 2010. Predators use environmental cues to discriminate between prey. *Behavioral Ecology and Sociobiology*, 64(12):1991–1997.
- Harlow, H. F. 1949. The formation of learning sets. *Psychological Review*, 56(1):51–65.
- Holland, P. C. and Schiffino, F. L. 2016. Mini-review: Prediction errors, attention and associative learning. *Neurobiology of Learning and Memory*, 131:207–215.
- Izquierdo, A., Brigman, J. L., Radke, A. K., Rudebeck, P. H., and Holmes, A. 2017. The neural basis of reversal learning: An updated perspective. *Neuroscience*, 345:12–26.
- Jacobs, R. A. 1988. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1(4):295–307.
- Janmaat, K. R., Boesch, C., Byrne, R., Chapman, C. A., Goné Bi, Z. B., Head, J. S., Robbins, M. M., Wrangham, R. W., and Polansky, L. 2016. Spatio-temporal complexity of chimpanzee food: How cognitive adaptations can counteract the ephemeral nature of ripe fruit. *American Journal of Primatology*, 78(6):626–645.
- Kazemi, B., Gamberale-Stille, G., Tullberg, B. S., and Leimar, O. 2014. Stimulus salience as an explanation for imperfect mimicry. *Current Biology*, 24(9):965–969.

- 495 Le Pelley, M. E. 2010. The hybrid modeling approach to conditioning. In Schmajuk, N.,  
 496 editor, *Computational Models of Conditioning*, pages 71–107. Cambridge University  
 497 Press, Cambridge, UK.
- 498 Lea, S. E. G., Chow, P. K. Y., Leaver, L. A., and McLaren, I. P. L. 2020. Behavioral  
 499 flexibility: A review, a model, and some exploratory tests. *Learning & Behavior*,  
 500 48(1):173–187.
- 501 Leimar, O., Dall, S. R. X., Houston, A. I., and McNamara, J. M. 2022. Behavioural  
 502 specialization and learning in social networks. *Proceedings of the Royal Society B:  
 503 Biological Sciences*, 289(1980):20220954.
- 504 Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., and Niv, Y. 2017. Dynamic  
 505 interaction between reinforcement learning and attention in multidimensional envi-  
 506 ronments. *Neuron*, 93(2):451–463.
- 507 Liu, Y., Day, L. B., Summers, K., and Burmeister, S. S. 2016. Learning to learn:  
 508 advanced behavioural flexibility in a poison frog. *Animal Behaviour*, 111:167–172.
- 509 Mackintosh, N. J. 1975. A theory of attention: Variations in the associability of stimuli  
 510 with reinforcement. *Psychological Review*, 82(4):276–298.
- 511 Mackintosh, N. J., Mcgonigle, B., and Holgate, V. 1968. Factors underlying improve-  
 512 ment in serial reversal learning. *Canadian Journal of Psychology*, 22(2):85–95.
- 513 Mahmood, A. R., Sutton, R. S., Degris, T., and Pilarski, P. M. 2012. Tuning-free  
 514 step-size adaptation. In *2012 IEEE International Conference on Acoustics, Speech  
 515 and Signal Processing (ICASSP)*, pages 2121–2124.
- 516 Miller, R. R., Barnet, R. C., and Grahame, N. J. 1995. Assessment of the Rescorla-  
 517 Wagner model. *Psychological Bulletin*, 117(3):363–386.
- 518 Nassar, M. R., Wilson, R. C., Heasley, B., and Gold, J. I. 2010. An approximately  
 519 Bayesian delta-rule model explains the dynamics of belief updating in a changing  
 520 environment. *Journal of Neuroscience*, 30(37):12366–12378.
- 521 Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., and  
 522 Wilson, R. C. 2015. Reinforcement learning in multidimensional environments relies  
 523 on attention mechanisms. *The Journal of Neuroscience*, 35(21):8145–8157.

- Pearce, J. M. and Hall, G. 1980. A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, 87(6):532–552.
- Pearce, J. M. and Mackintosh, N. J. 2010. Two theories of attention: a review and possible integration. In Mitchell, C. J. and Le Pelley, M. E., editors, *Attention and Associative Learning: From Brain to Behaviour*, pages 11–40. Oxford University Press, Oxford, UK.
- Pierce, B. J. and McWilliams, S. R. 2005. Seasonal changes in composition of lipid stores in migratory birds: causes and consequences. *The Condor*, 107(2):269–279.
- Piray, P. and Daw, N. D. 2021. A model for learning based on the joint estimation of stochasticity and volatility. *Nature Communications*, 12(1):6587.
- Rescorla, R. A. and Wagner, A. R. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical conditioning II: current research and theory*, pages 64–99. Appleton-Century-Crofts, New York.
- Rmus, M., McDougale, S. D., and Collins, A. G. 2021. The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, 38:66–73.
- Sol, D., Duncan, R. P., Blackburn, T. M., Cassey, P., and Lefebvre, L. 2005. Big brains, enhanced cognition, and response of birds to novel environments. *Proceedings of the National Academy of Sciences*, 102(15):5460–5465.
- Sol, D., Timmermans, S., and Lefebvre, L. 2002. Behavioural flexibility and invasion success in birds. *Animal Behaviour*, 63(3):495–502.
- Soltani, A. and Izquierdo, A. 2019. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, 20(10):635–644.
- Sutton, R. S. 1992a. Adapting bias by gradient descent: An incremental version of delta-bar-delta. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 171–176. MIT Press, Cambridge, MA.
- Sutton, R. S. 1992b. Gain adaptation beats least squares? In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, pages 161–166. Yale University, New Haven, CT.

- 554 Sutton, R. S. 2022. A history of meta-gradient: Gradient methods for meta-learning.  
555 *arXiv preprint, 2202.09701*, pages 1–8.
- 556 Sutton, R. S. and Barto, A. G. 2018. *Reinforcement learning: An introduction second*  
557 *edition*. MIT Press, Cambridge, MA.
- 558 Szabo, B., Damas-Moreira, I., and Whiting, M. J. 2020. Can cognitive ability give  
559 invasive species the means to succeed? A review of the evidence. *Frontiers in*  
560 *Ecology and Evolution*, 8:187.
- 561 Torrents-Rodas, D., Koenig, S., Uengoer, M., and Lachnit, H. 2021. Evidence for  
562 two attentional mechanisms during learning. *Quarterly Journal of Experimental*  
563 *Psychology*, 74(12):2112–2123.
- 564 Triki, Z., Granell-Ruiz, M., Fong, S., Amcoff, M., and Kolm, N. 2022. Brain morphology  
565 correlates of learning and cognitive flexibility in a fish species (*Poecilia reticulata*).  
566 *Proceedings of the Royal Society B: Biological Sciences*, 289(1978):20220844.
- 567 Triki, Z., Wismer, S., Rey, O., Ann Binning, S., Levorato, E., and Bshary, R. 2019.  
568 Biological market effects predict cleaner fish strategic sophistication. *Behavioral*  
569 *Ecology*, 30(6):1548–1557.
- 570 Trimmer, P. C., McNamara, J. M., Houston, A. I., and Marshall, J. A. R. 2012. Does  
571 natural selection favour the Rescorla–Wagner rule? *Journal of Theoretical Biology*,  
572 302:39–52.
- 573 Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z.,  
574 Hassabis, D., and Botvinick, M. 2018. Prefrontal cortex as a meta-reinforcement  
575 learning system. *Nature Neuroscience*, 21(6):860–868.
- 576 Warren, J. M. 1966. Reversal learning and the formation of learning sets by cats and  
577 rhesus monkeys. *Journal of Comparative and Physiological Psychology*, 61(3):421–  
578 428.
- 579 Wilson, B., Mackintosh, N. J., and Boakes, R. A. 1985. Transfer of relational rules  
580 in matching and oddity learning by pigeons and corvids. *The Quarterly Journal of*  
581 *Experimental Psychology Section B*, 37(4b):313–332.
- 582 Wismer, S., Grutter, A., and Bshary, R. 2016. Generalized rule application in  
583 bluestreak cleaner wrasse (*Labroides dimidiatus*): using predator species as social  
584 tools to reduce punishment. *Animal Cognition*, 19(4):769–778.

# Figures

585

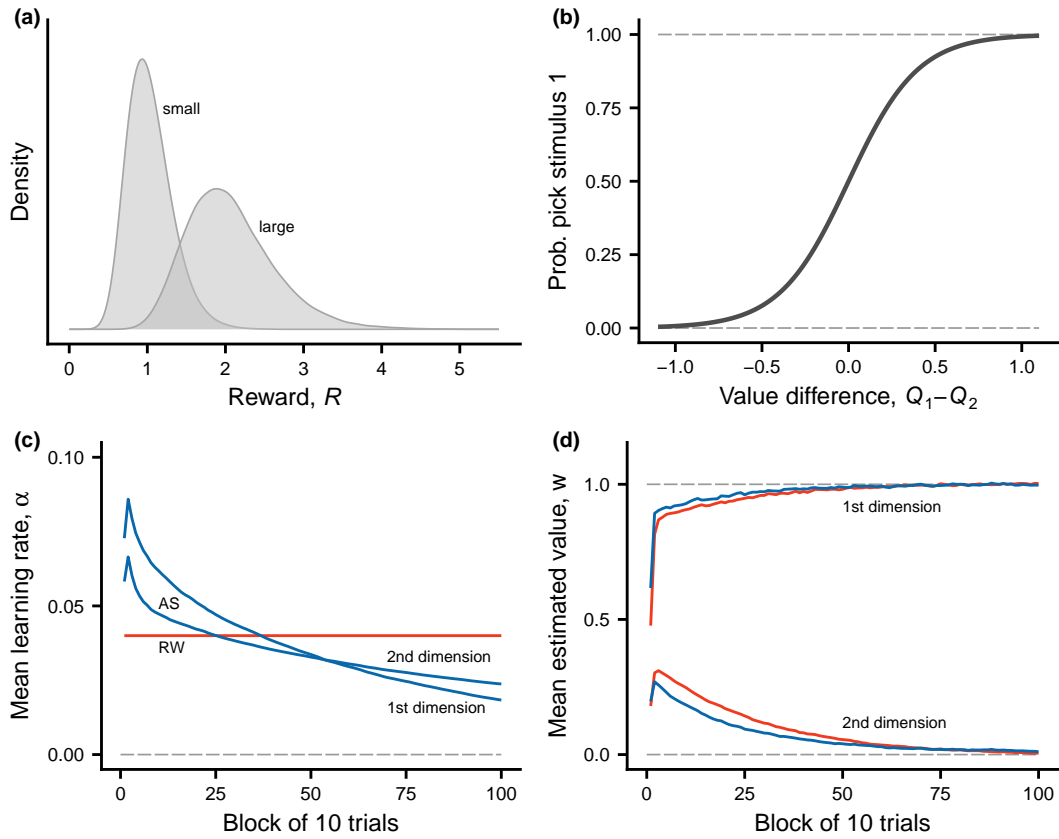


Figure 1: Overview of the first phase, where individuals learn to discriminate between two types of compound stimuli ('small' and 'large' clients). **(a)** Distribution of rewards from the two types of compound stimuli. **(b)** The function from equation (4), giving the probability of choice from the difference in estimated values of the two compound stimuli present in a trial. **(c)** Learning rates for Rescorla-Wagner (RW) and Autostep (AS) for the two stimulus dimensions. **(d)** Estimated values for Rescorla-Wagner and Autostep for the two stimulus dimensions (first dimension has true value 1.0 and second has true value 0). There are 10 trials in a block and data are averages over 100 replicate learning simulations.

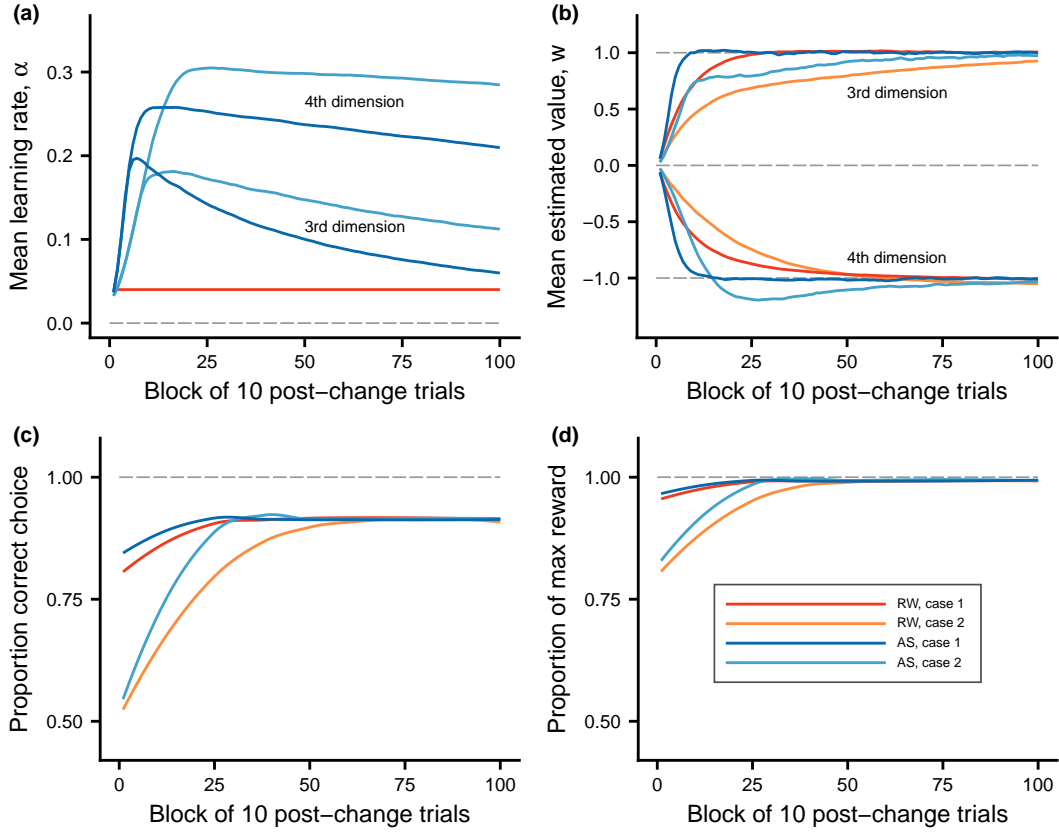


Figure 2: Comparisons of the second phase of learning, when the world becomes more complex, between Rescorla-Wagner (RW) and Autostep (AS), and for the two cases studied. Colour coding in panel (d) applies to all panels. (a) Learning rates for the different learning algorithms and cases (note that the learning rate for Rescorla-Wagner is constant). As an illustration, the third and fourth stimulus dimensions are shown. Note that the features in these dimensions were not present in the first phase. The results are similar for the other new dimensions in Table 1 (dimensions 5-10). (b) Estimated values for the different learning algorithms and cases, for stimulus dimensions 3 and 4. (c) Proportion of choices that are correct, in the sense of the individual choosing the compound stimulus with higher true value. (d) Proportion of reward gained out of the maximum true expected reward available in a trial.

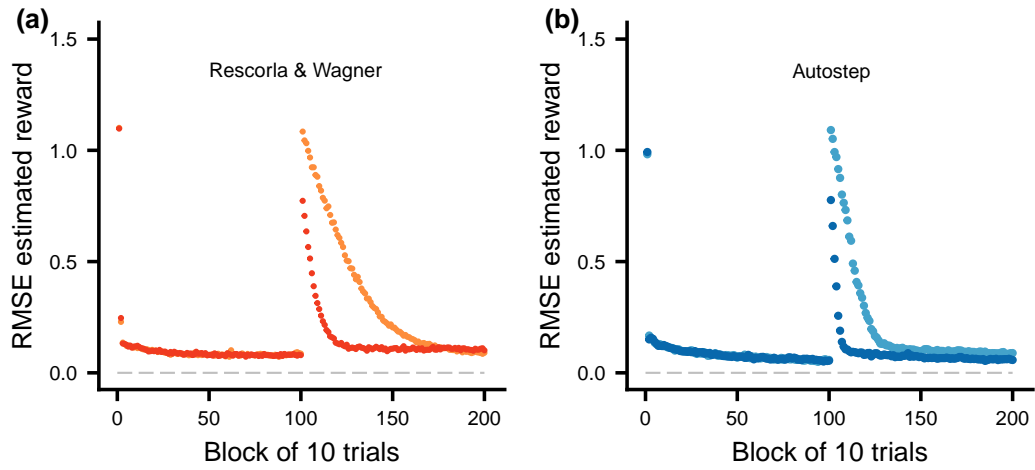


Figure 3: Illustration of the root mean square error (RMSE) of the individual's estimate ( $Q$ ) of the reward from the selected compound stimulus, plotted against the trial block, over both phases of learning. There are 10 trials in a block and data are averages over 100 replicate learning simulations. (RMSE is similar to a standard deviation but instead measuring the deviation of an estimate from the true value.) **(a)** Rescorla-Wagner learning, with  $\alpha_{RW} = 0.04$ . **(b)** Autostep learning, following Mahmood et al. (2012). The colour coding is as in Figure 2d.



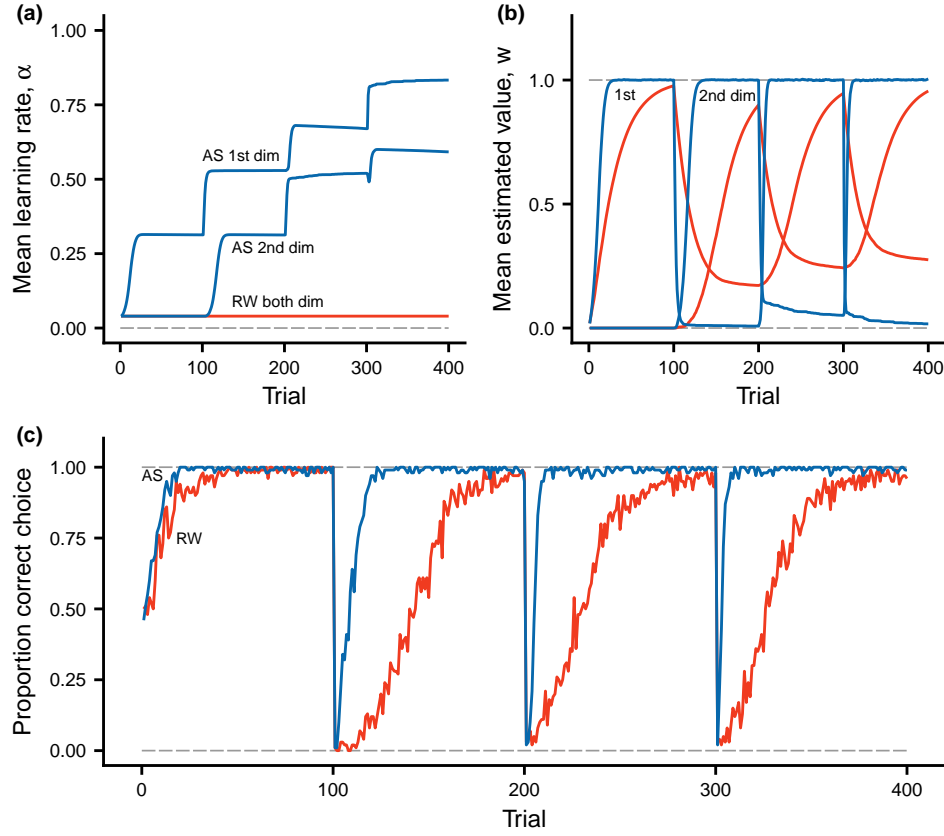


Figure 4: Reversal learning simulation. There are two stimulus dimensions, with 0/1 features that each indicate a type of stimulus. In the first phase of 100 trials, choosing the stimulus with a feature in dimension 1 is rewarded ( $R = 1$ ) and choosing the other, with a feature in dimension 2, is not rewarded ( $R = 0$ ). In the next 100 trials, the rewards are reversed, and then the procedure is repeated for another 200 trials. **(a)** Learning rates for the Rescorla-Wagner (RW) and Autostep (AS) algorithms (note that the learning rate for Rescorla-Wagner is constant). **(b)** Estimated values for RW and AS, for the two stimulus dimensions. **(c)** Proportion of choices that are correct, in the sense of the individual choosing the stimulus with higher true value.

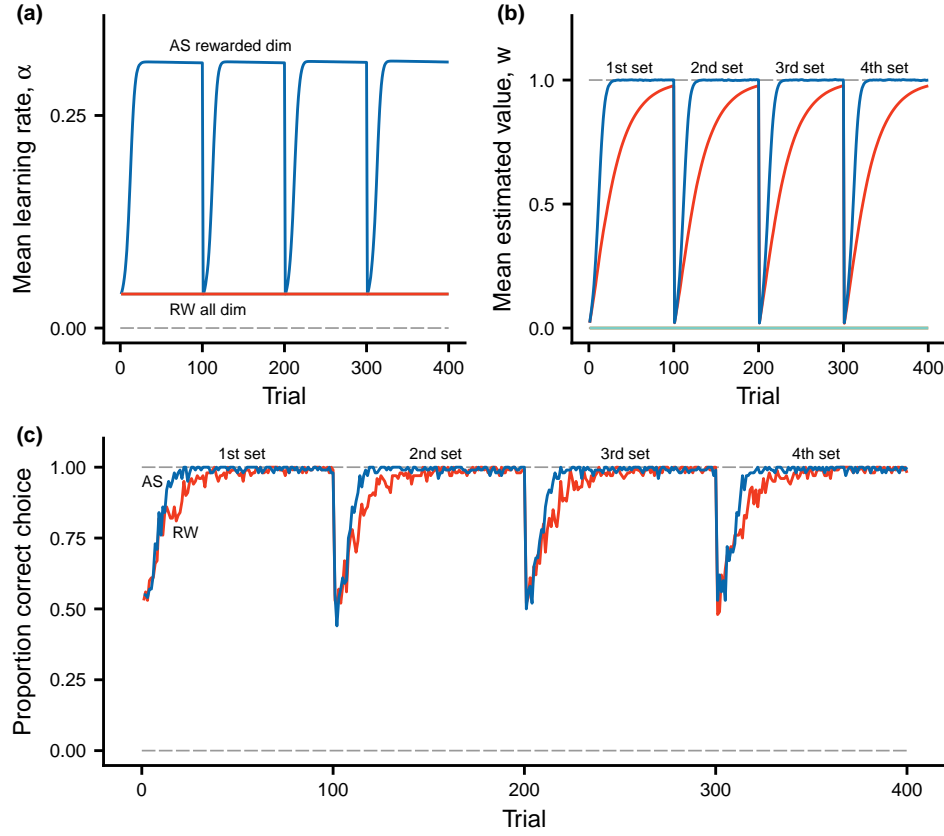


Figure 5: Learning set formation. There are 8 stimulus dimensions, with 0/1 features that each indicate a type of stimulus. In the first phase of 100 trials, there is a choice between the first pair of stimuli, where one is rewarded ( $R = 1$ ) and the other is unrewarded ( $R = 0$ ). In the next 100 trials, a new pair of rewarded and unrewarded is used, and so on until four pairs have been used. **(a)** Learning rates for the Rescorla-Wagner (RW) and Autostep (AS) algorithms (note that the learning rate for Rescorla-Wagner is constant). **(b)** Estimated values for RW and AS, for the rewarded stimulus dimension in each set, and for the unrewarded dimension (green curve at bottom). **(c)** Proportion of choices that are correct, in the sense of the individual choosing the stimulus with higher true value.

# Tables

Table 1: Characteristics of stimulus dimensions and compound stimuli used for the simulation of a change to a more complex world. There are 10 compound stimuli (CS1 to CS10) that can be distinguished using 10 stimulus dimensions. The first dimension represents size, with expected values small (1.0) and large (2.0), and the others are absence/presence (0/1) dimensions. The expected reward values per feature ( $W_m$ ) are given in the second column and the features of the different compound stimuli are in the following columns.

[illegible]

## 587 Supporting information

### 588 Additional figures

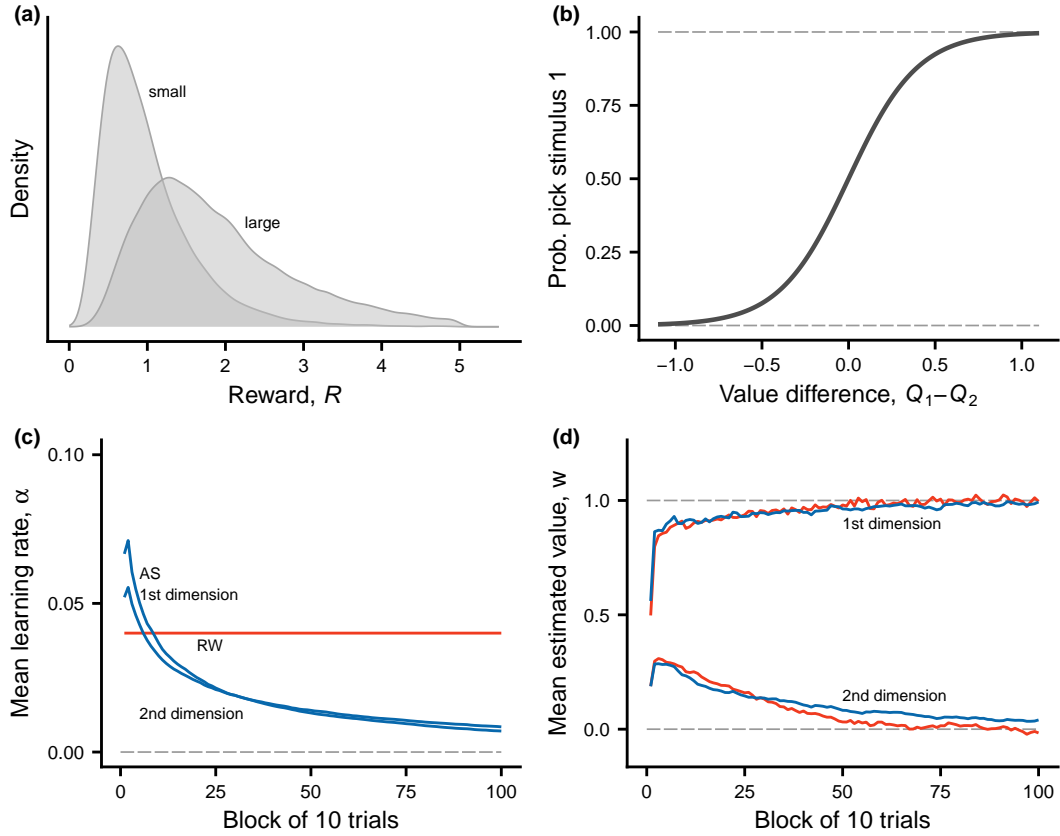


Figure S1: Same as the first cases in the main text (Figure 1), but with high reward stochasticity ( $\sigma_R = 0.50$ ). (a) Distribution of rewards from the two types of compound stimuli. (b) The function from equation (4), giving the probability of choice from the difference in estimated values of the two compound stimuli present in a trial. (c) Learning rates for Rescorla-Wagner (RW) and Autostep (AS) for the two stimulus dimensions. (d) Estimated values for Rescorla-Wagner and Autostep for the two stimulus dimensions (first dimension has true value 1.0 and second has true value 0). There are 10 trials in a block and data are averages over 100 replicate learning simulations.

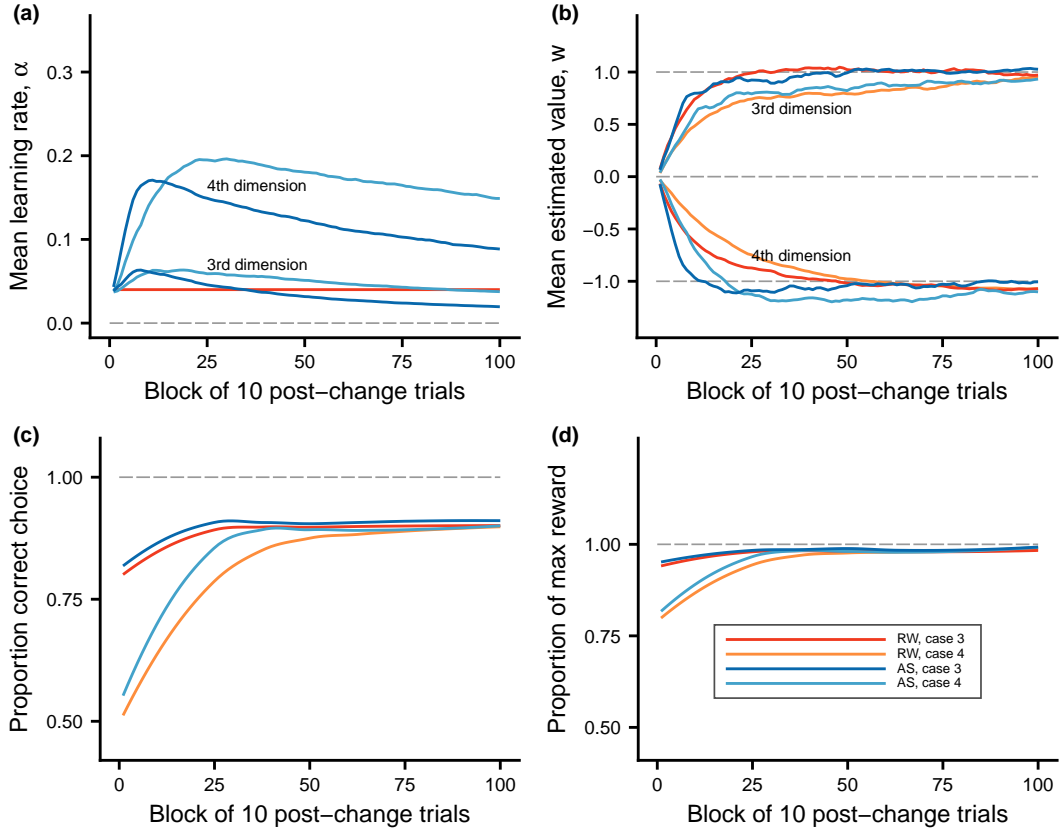


Figure S2: Same as the first cases in the main text (Figure 2), but with high reward stochasticity ( $\sigma_R = 0.50$ ). Comparisons of the second phase of learning, when the world becomes more complex, between Rescorla-Wagner (RW) and Autostep (AS), and for the cases studied. Colour coding in panel (d) applies to all panels. (a) Learning rates for the different learning algorithms and cases (note that the learning rate for Rescorla-Wagner is constant). As an illustration, the third and fourth stimulus dimensions are shown. Note that the features in these dimensions were not present in the first phase. (b) Estimated values for the different learning algorithms and cases, for stimulus dimensions 3 and 4. (c) Proportion of choices that are correct, in the sense of the individual choosing the compound stimulus with higher true value. (d) Proportion of reward gained out of the maximum true expected reward available in a trial.

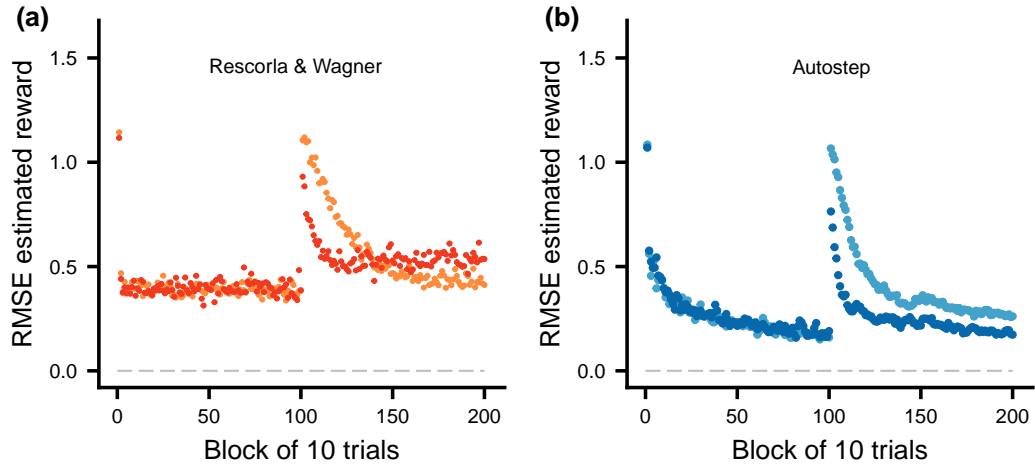


Figure S3: Same as the first cases in the main text (Figure 3), but with high reward stochasticity ( $\sigma_R = 0.50$ ). Illustration of the root mean square error (RMSE) of the individual's estimate ( $Q$ ) of the reward from the selected compound stimulus, plotted against the trial block, over both phases of learning. There are 10 trials in a block and data are averages over 100 replicate learning simulations. **(a)** Rescorla-Wagner learning, with  $\alpha_{RW} = 0.04$ . **(b)** Autostep learning, following Mahmood et al. (2012). The colour coding is as in Figure S2d.

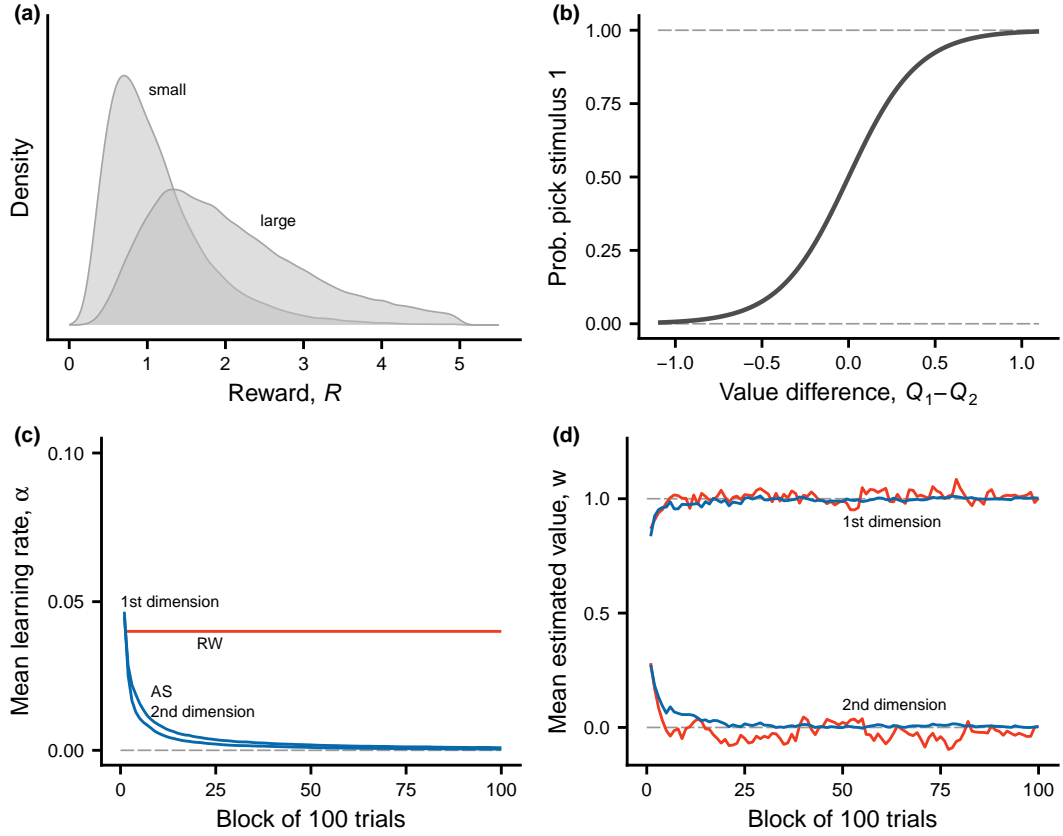


Figure S4: High reward stochasticity ( $\sigma_R = 0.50$ ), as in Figure S1, but the first phase of learning is much longer, with  $T = 10000$ . **(a)** Distribution of rewards from the two types of compound stimuli. **(b)** The function from equation (4), giving the probability of choice from the difference in estimated values of the two compound stimuli present in a trial. **(c)** Learning rates for Rescorla-Wagner (RW) and Autostep (AS) for the two stimulus dimensions. **(d)** Estimated values for Rescorla-Wagner and Autostep for the two stimulus dimensions (first dimension has true value 1.0 and second has true value 0). There are 100 trials in a block and data are averages over 10 replicate learning simulations.



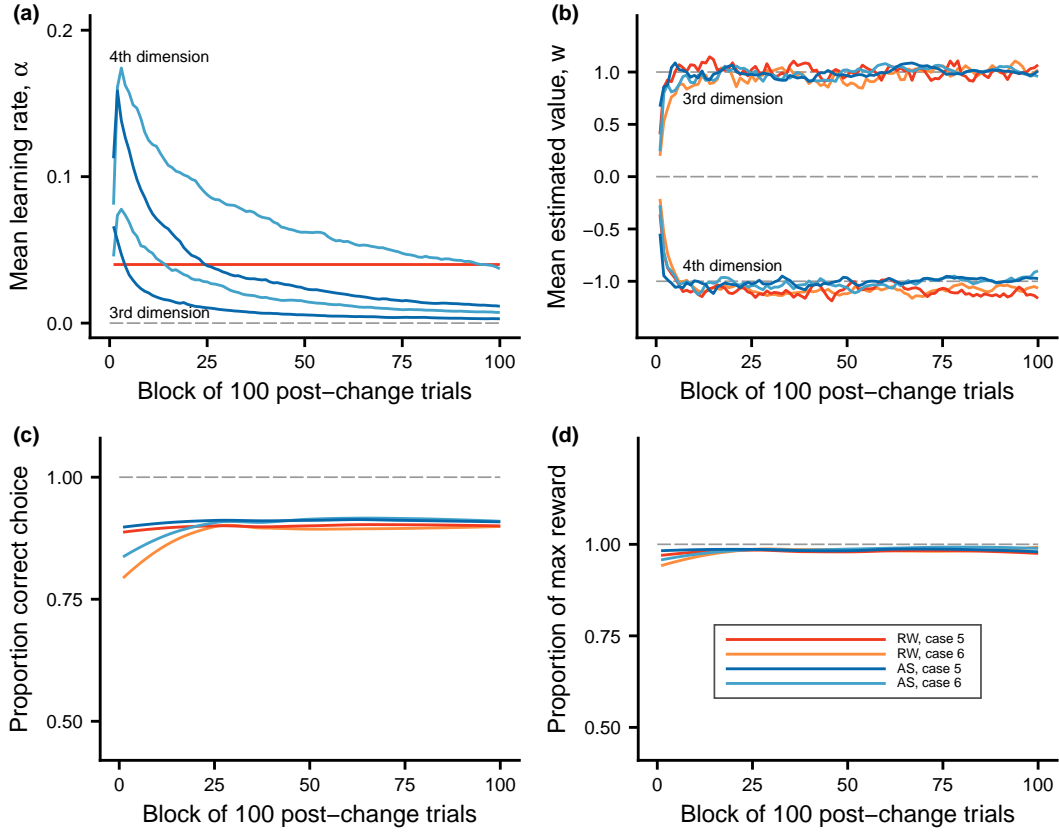


Figure S5: High reward stochasticity ( $\sigma_R = 0.50$ ), as in Figure S2, but the second phase of learning is much longer, with  $T = 10000$ . Comparisons of the second phase of learning, when the world becomes more complex, between Rescorla-Wagner (RW) and Autostep (AS), and for the cases studied. Colour coding in panel (d) applies to all panels. **(a)** Learning rates for the different learning algorithms and cases (note that the learning rate for Rescorla-Wagner is constant). As an illustration, the third and fourth stimulus dimensions are shown. Note that the features in these dimensions were not present in the first phase. **(b)** Estimated values for the different learning algorithms and cases, for stimulus dimensions 3 and 4. **(c)** Proportion of choices that are correct, in the sense of the individual choosing the compound stimulus with higher true value. **(d)** Proportion of reward gained out of the maximum true expected reward available in a trial.

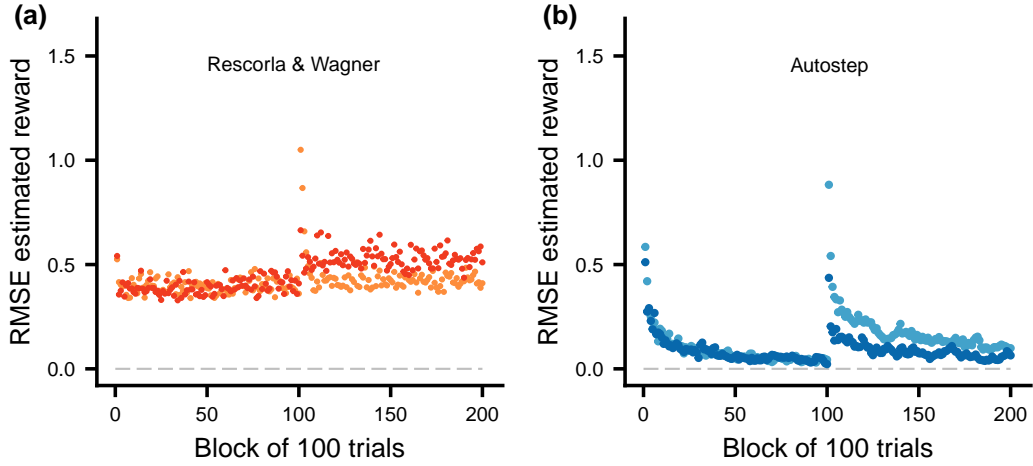


Figure S6: High reward stochasticity ( $\sigma_R = 0.50$ ), as in Figure S3, but the first and second phases of learning are much longer, with  $T = 10000$ . Illustration of the root mean square error (RMSE) of the individual’s estimate ( $Q$ ) of the reward from the selected compound stimulus, plotted against the trial block, over both phases of learning. There are 100 trials in a block and data are averages over 10 replicate learning simulations. **(a)** Rescorla-Wagner learning, with  $\alpha_{RW} = 0.04$ . **(b)** Autostep learning, following Mahmood et al. (2012). The colour coding is as in Figure S5d.

## Model details

Using the notation  $W_m$  for the true expected value for a stimulus dimension (e.g., as given in Table 1 of the main text), the true expected reward from a compound stimulus with stimulus components or features  $x_m$ ,  $m = 1, \dots, M$ , is given by

$$\bar{R} = \sum_{m=1}^M W_m x_m. \quad (\text{S1})$$

For random variation in the rewards, one possibility is to assume additive variation, but we instead assume that  $R$  is log-normally distributed, preventing negative values. Thus,

$$R = \exp(\mu_R + z_R), \quad (\text{S2})$$

with  $\mu_R$  such that  $\bar{R} = \exp(\mu_R)$ , i.e.  $\mu_R = \log(\bar{R})$ , where  $z_R$  is normally distributed with mean zero and standard deviation  $\sigma_R$ . We might, for instance, have  $\sigma_R = 0.1$ , which is used in Figures 1 to 3 in the main text.

Our learning models, as described by equations (1, 2, 3) in the main text and equations (S1, S2) here, are examples of action-value learning. One can view such

learning as a modification of classical conditioning, making it applicable to instrumental conditioning (see sections 2.2 and 2.5 in Sutton and Barto (2018) for discussion of this learning approach). Note also that action-value learning can be regarded as a simplified version of the Sarsa algorithm, for cases where individuals do not use any sophisticated states and where each learning trial is a separate episode (terminology from Sutton and Barto (2018)). We get a connection to the presentation in Sutton and Barto (2018) by assuming that the state in a trial is just the compound stimuli that are present in that trial, for the individual for choose between.

## Learning rates

### Rescorla-Wagner

The Rescorla-Wagner learning mechanism has learning rates that are constant in time. In our simulations, we assume that

$$\alpha_{mt} = \alpha_{RW} \quad (\text{S3})$$

for the learning rate for dimension  $m$  in trial  $t$ . Thus, we assume that the Rescorla-Wagner learning rate  $\alpha_{RW}$  is constant, and thus independent of the stimulus dimension  $m$  and the trial  $t$  (we use  $\alpha_{RW} = 0.04$  in our simulations).

### IDBD

The IDBD (Incremental Delta-Bar-Delta) learning was developed by Sutton (1992a), and is also described in Sutton (1992b) and by equations (1) and (3) in Mahmood et al. (2012). For clarity we write the prediction error from equation (2) in trial  $t$  as

$$\delta_t = R_t - Q_t, \quad (\text{S4})$$

where  $Q_t$  is the estimated value from equation (3) for the chosen compound stimulus with stimulus components  $x_m$ . We write the estimated value as

$$Q_t = \sum_{m=1}^M x_m w_{m,t}. \quad (\text{S5})$$

Using the order of updates for IDBD from Sutton (1992a,b) and Mahmood et al. (2012), there are initial learning rates  $\alpha_{m,1}$  and, to update the estimated values, we first update

the learning rates as follows:

$$\alpha_{m,t+1} = \alpha_{m,t} \exp(\mu \delta_t x_m h_{m,t}), \quad (\text{S6})$$

where  $\mu$  is a meta learning rate,  $\delta_t$  is from equation (S4), and  $h_{m,t}$  is an additional quantity with starting value  $h_{m,1} = 0$ . The estimated value updates, corresponding to equation (1), are then

$$w_{m,t+1} = w_{m,t} + \alpha_{m,t+1} x_m \delta_t. \quad (\text{S7})$$

The quantity  $h_{m,t}$  is updated as follows

$$h_{m,t+1} = h_{m,t} [1 - \alpha_{m,t+1} x_m^2]^+ + \alpha_{m,t+1} x_m \delta_t. \quad (\text{S8})$$

The notation  $[X]^+$  means equal to  $X$  for positive  $X$  and zero otherwise.

For an intuitive interpretation, note the quantity  $h_m$  is a kind of memory or trace of changes to  $w_m$ , because those changes are given by  $\alpha_{m,t+1} x_m \delta_t$ . The exponent in equation (S6) contains the product of this trace with  $\delta_t x_m$ , which is proportional to the current change in  $w_m$ . Thus, if successive changes to  $w_m$  tend to be positively correlated, the learning rate  $\alpha_m$  will increase, and similarly decrease for negative correlations.

## Autostep

A problem with the IDBD algorithm, pointed out by Mahmood et al. (2012), is that it is very sensitive to the exact value of the meta learning rate  $\mu$ . To avoid this problem, Mahmood et al. (2012) introduced changes and obtained a more robust algorithm, which they called Autostep. It is called Autostep because the step size, or effective meta-learning rate, is automatically adjusted to an appropriate value. The Autostep algorithm is given in Table 1 of Mahmood et al. (2012), and is as follows.

First set the two meta learning parameters  $\mu$  and  $\tau$  (we used  $\mu = 0.2$  and  $\tau = 100$  in our Autostep simulations). Then initialise  $w_{m,1}$  and  $\alpha_{m,1}$  (we used  $w_{m,1} = 0$  and  $\alpha_{m,1} = 0.04$ ) and set  $v_{m,1} = 0$  and  $h_{m,1} = 0$ .

For each trial  $t$  we then have the following. The estimated values for compound stimuli are computed from features  $x_m$  as

$$Q_t = \sum_{m=1}^M x_m w_{m,t}, \quad (\text{S9})$$

just as previously. The learner chooses a compound stimulus using the soft-max rule in equation (4) and perceives a reward  $R$ . This gives a prediction error

$$\delta_t = R_t - Q_t, \quad (\text{S10})$$

just as previously. Then compute the quantity

$$v_{m,t+1} = \max(|\delta_t x_m h_{m,t}|, v_{m,t} + \frac{1}{\tau} \alpha_{m,t} x_m^2 (|\delta_t x_m h_{m,t}| - v_{m,t})). \quad (\text{S11})$$

If  $v_{m,t+1} \neq 0$ , compute

$$\alpha_{m,\text{temp}} = \alpha_{m,t} \exp\left(\frac{\mu \delta_t x_m h_{m,t}}{v_{m,t+1}}\right), \quad (\text{S12})$$

otherwise use  $\alpha_{m,\text{temp}} = \alpha_{m,t}$ . Do this for each  $m$  and compute

$$S_\alpha = \max\left(\sum_{m=1}^M \alpha_{m,\text{temp}} x_m^2, 1\right). \quad (\text{S13})$$

Use this  $S_\alpha$  to normalise the learning rates

$$\alpha_{m,t+1} = \frac{\alpha_{m,\text{temp}}}{S_\alpha}. \quad (\text{S14})$$

These are then used for the updates of the estimated values:

$$w_{m,t+1} = w_{m,t} + \alpha_{m,t+1} x_m \delta_t. \quad (\text{S15})$$

Finally, the  $h_{m,t}$  are updated as follows:

$$h_{m,t+1} = h_{m,t} [1 - \alpha_{m,t+1} x_m^2] + \alpha_{m,t+1} x_m \delta_t. \quad (\text{S16})$$

The changes in Autostep from IDBD are, first, that the exponent in the update in equation (S6) is divided by the quantity  $v_{m,t+1}$  in equation (S12), achieving a reasonable effective meta step size and, second, that the learning rates are normalised in equation (S14). See Mahmood et al. (2012) for more comments on the Autostep algorithm.