Freedom from habits: the capacity for autonomous behaviour

Keiji Ota[1], Lucie Charles[1], Patrick Haggard[1]

[1] Institute of Cognitive Neuroscience, University College London, London, United Kingdom

Corresponding author

Keiji Ota

Keywords

Cognitive control / Free action / Habits / Volition / Competitive pressure

Abstract

The capacity for autonomous behaviour is key to human intelligence, and fundamental to modern social life. However, experimental investigations of the cognitive bases of human autonomy are challenging, because experimental paradigms typically constrain behaviour using controlled contexts, and elicit behaviour by external triggers. In contrast, the sources of human autonomy and freedom are assumed to be endogenous. Here we propose a new theoretical construct of adaptive autonomy, meaning the capacity to make behavioural choices that are free from constraints of both immediate triggers and habitual responding. Participants played a competitive game in which they had to choose the right time to act, in the face of an opponent who punished (in separate blocks) either choice biases, habitual sequences of action timing across trials, or habitual responses to the effects of reinforcement. Adaptive autonomy with respect to each habit was measured by the ability to maintain performance against the opponent even when the corresponding habit was punished. We found that participants were able, under pressure from their opponent, to become free from habitual choices of when to act, but were not able to free themselves from win-stay, lose-shift patterns of reinforcement, even when these resulted in punishment. These results propose a new testing ground of autonomous behaviour as a flexible adaptation of more or less habitual behaviours that co-exist with different classes of external constraint.

Introduction

Animal behaviour depends upon both exogenous, environmental factors and endogenous factors. The endogenous factors can be conceptualized as a dimension extending from stereotypical behavioural patterns like habits, and flexible, intelligent actions. The latter are thought to play a special role in human autonomy and volition. Adaptability and variation of behavioural choices allows humans to adapt to environmental challenges and find novel solutions. The capacity of human autonomy has been extensively examined in experimental tasks which encourage participants to act freely (Brass & Haggard, 2007; Fleming et al., 2009; Jahanshahi et al., 1995; Libet et al., 1983) or to act randomly (Baddeley et al., 1998; Baddeley, 1966; Jahanshahi et al., 2000) by explicitly telling them to do so. Such instructions in voluntary-action studies invite participants to behave in a way that reflects their understanding of volition and freedom (Haggard, 2008). Outside the laboratory, in contrast, people readily switch between more stereotyped and more autonomous behaviours without explicit instruction, as a function of multiple situational and internal factors.

As such, it remains unclear the extent to which people can express their behavioural autonomy through volitional actions. Competitive games might be a good testing ground for this question, for two reasons. First, many competitive games require people to initiate an action endogenously, rather than in response to an external stimulus. For example, in the ‘rock, paper, scissors’ game, each participant selects an action without first seeing the action of their opponent. This stimulus-independence is considered a necessary condition for volition (Jenkins et al., 2000). Second, volitional actions are often contrasted with habitual, or routine actions (Haggard, 2019). Competitive games offer a convenient way to manipulate the extent to which any individual action is or is not habitual. For example, if a player behaves habitually in a competitive game, their opponent will be able to predict their upcoming choice, and adjust their strategy accordingly. Therefore, an agent playing a competitive game should avoid habitual or exploitative behaviours, and must innovate in order to avoid being predicted. Non-human primates indeed respond to competitive pressure by initiating exploratory behaviour (Barraclough et al., 2004; Lee et al., 2004; Lee et al., 2005).

The present study therefore investigates human volition and autonomy in a competitive game task in which participants could not react to their competitor’s current move (stimulus independence), and additionally could receive reward only when they avoided the competitor’s prediction (habit independence). Further, we used several virtual competitor algorithms, each one designed to punish a particular kind of habit. Whereas psychologists have often thought of habits as personality traits, we reasoned that a person may stop behaving habitually when a competitor begins to predict, exploit and punish their habitual beahviour. Thus, the *change* in habitual behaviour under competitive pressure offers a quantitative measure of individual autonomy with respect to habits.

We conceptualised three distinct “habitual families”[1]. The first was automatic response selection (Dolan & Dayan, 2013; Du et al., 2022; Robbins & Costa, 2017). We will refer to this as *standard choice habits*. Consider the simple task of generating one of three digits in each turn (see Figure 1A). Agent X may prefer to choose “1”, for whatever reason, while agent Y may be less biased. The statistical similarity between an individual’s observed choice pattern and a random pattern can be measured (depicted by the right arrow in Figure 1A). Suppose now that a competitor punishes X for repeating one choice within a game scenario. If X can break her habit, she should now choose the two other digits more often (shared area in Figure 1A). Agent Y may be less adaptive and stick to his original choice pattern. This adaptive capacity may reflect the autonomy each agent has over their choice habits (left arrow in Figure 1A). We call this quantity *adaptive autonomy*.

The second habit family we considered is *transition habits*. This refers to action chains or routines (Lashley, 1951; Robbins & Costa, 2017; Rosenbaum et al., 2007). In our example, integer counting (“1, 2, 3”) is such a habit (agent X in Figure 1B). The only way to completely avoid such habits is to generate each choice independently from the previous trial. Yet studies of random number generation show people find this difficult (Baddeley et al., 1998; Baddeley, 1966; Bar-Hillel & Wagenaar, 1991). The shaded area and left arrow in Figure 1B illustrate potential behavioural adaptation to punishing transition habits.

Lastly, we address how people respond to action successes and failures, by

[1] The concept of a habit has been discussed over centuries by philosophers (Barandiaran & Di Paolo, 2014). Its theoretical framework has been formalised by psychologists and neuroscientists’ works, but it is still controversial (Dolan & Dayan, 2013; Du et al., 2022; Robbins & Costa, 2017). A psychologist's view of a habit is that it reflects the formation of stimulus-response associations (Wood & Runger, 2016). A traditional testing ground for habits is a reward devaluation paradigm. Here a reward previously assigned to a stimulus-response association is devalued (Robbins & Costa, 2017). If the response is still automatically evoked by the stimulus, its behaviour is said to be a habit rather than goal-directed. In the present study, we do not aim to examine whether or how habits are formed. Rather, we investigate how people become liberated from their endogenous habitual patterns. The three habit families we conceptualised are considered different expressions of habits. From the perspective of a stimulus-response association, standard choice habits are any obvious responses in any task. Rule habits are those responses evoked by the last response. While reinforcement habits are the responses elicited by the last outcome/feedback. The precise mechanism by which each habit forms is unimportant. Rather, the concept of three habit families refers to statistical patterns that are not random nor independent (see below) and that may be continually shaped by ongoing experience. Thus, volition in our paradigm requires the regulation of salient/habitual patterns and the exploration of new behavioural patterns.

considering *reinforcement habits*. In Figure 1C, agent X show typical win-stay lose-shift behaviour in a digit generation task. In contrast, a habit-free agent generates each choice independently from whether the previous outcome was rewarded or not. The vast majority of studies in reinforcement learning assume that a 'win-stay lose-shift' strategy is natural, or even unavoidable (Worthy et al., 2013). Here we test whether people can unlearn this familiar reinforcement habit when it is punished by a competitor. Adaptive autonomy would mean that an agent would be able to break the association between their next action and the previous outcome (see the potential change in Fig. 1C).

In this experiment, we designed a structured series of competitors in a game scenario, in order to selectively punish these specific habits, and measure individuals' capacity for adaptive autonomy, as the change in behaviour when a specific habit family was punished. Using this framework, we tested the capacities or limits of human autonomy for three "habitual families" of choice, transition and reinforcement. To explore whether adaptive autonomy reflects a general capacity, or rather is specific to a particular habit family, we explored correlations across individuals in our adaptive autonomy measures for each habit. Finally, we modelled the learning process by which people generated a new action in order to avoid the competitor.

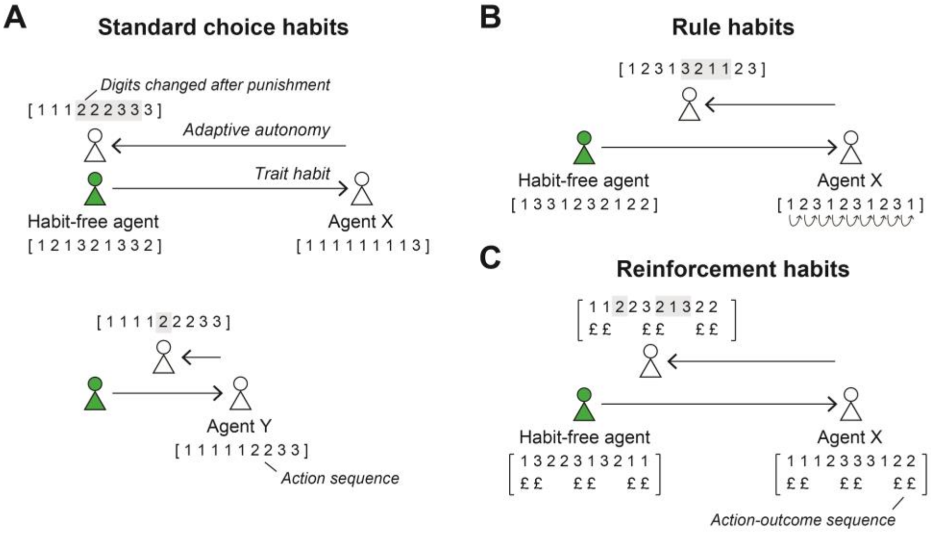


Figure 1. Three habitual families in a hypothetical experiment. An agent is asked to generate one digit from three in each turn. A sequence of generated digits is shown in a square bracket from left to right. The precise task is not important. **A**. Agent X habitually selects the digit "1" while agent Y occasionally selects the other two digits. A habit-free agent should select each of the 3 digits randomly. A right arrow indicates a pattern similarity from the habit-free agent to

agent X or Y. **B.** Agent X has a trait rule-based transition habit, in which they count up from the last digit. The habit-free agent selects the digit independently from the previous digit. **C.** In certain situations where an agent is rewarded, the pattern that is dependent on the reward assigned captures a trait reinforcement habit. In all habitual families, we measure the extent to which selection patterns change when agents are punished for habitual action. Shaded areas represent digits changed after the punishment and hypothetical agents move their location closer to the habit-free agent. A left arrow illustrates adaptive autonomy, a change in behavioural patterns with respect to habits.

Results

Experimental task

Participants were asked to decide when to press a key that caused some food to be delivered to a storage location. They were competing with a virtual competitor, represented as a flock of birds (Fig. 2A). The birds tried to catch the food during the delivery process, by deciding when to fly out of a tree and across the field. The participant's task was to deliver the food without it being caught by the birds. We programmed the birds to predict the time of the participant's next action based on the history of their reaction times. Based on this prediction, the birds made a choice of when to fly on each trial. They flew at a time that was designed to intercept the food thrown by a participant within one of three intervals: 1) early throw (0–1.5 sec), 2) middle throw (1.5–3.0 sec) or 3) late throw (3.0–4.5 sec). The participants could win a trial by pressing the key during one of two intervals that the birds did not select. The intervals were not explicitly demarcated for the participant, who experienced a continuum of potential action times in each trial. There was no time for participants to perform the task reactively because the birds could travel much faster than the food. If the participants simply waited for a moment when no birds flew and then threw, the birds could suddenly appear and intercept the food. Therefore, the participants were asked to predict when the birds would appear and avoid them. This feature means that our participant's actions were stimulus-independent.

There were 4 blocks in total. In each block, the participants competed with a class of competitor that pressurised a specific habit (Fig. 2B). In the baseline block, Competitor 0 was programmed to punish the participants for being impatient: the birds consistently punished a participant who threw in the early interval, so the participant was incentivized to wait to avoid being intercepted. In block 1, Competitor 1 punished standard choice habits, if a participant selected one interval more often than the other two. In block 2, Competitor 2 predicted transition habits, and punished any association between the time of the participant's current throw and the time of the preceding throw. Finally, in block 3, Competitor 3 punished reinforcement habits by seeking out whether the time of the current throw was associated with both the time of the preceding throw and the preceding outcome. Thus, the participants played against competitors who had increasingly sophisticated predictive power in each block. The participants required progressive degrees of autonomy across blocks: they needed to act in a way that was even more unconstrained than required by the competitors they had played previously.

Participants did not receive any explicit instruction or explanation about what habits they should avoid. The participants were never told when they should act on any given trial. Instead, they could only monitor the success/failure of avoiding the birds on each trial, and adapt their behaviour accordingly to try to avoid the birds on future trials. Thus, successful performance under different punishment regimes would depend on implicit mechanisms of adaptation rather than explicit understanding.

We first examined whether the predictive power of our protocol increased by checking the percentage of successful bird-avoiding trials. The participants achieved near perfect success rates against Competitor 0 who punished impatience (Fig. 2C; Median [Mdn] = 96.6%). The participants avoided an immediate response and initiated the throw 1.5 seconds after the trial starts on almost all trials (Fig. 2D upper panel). In block 1, the success rate dropped to 66.6%, as would be expected from purely stochastic choices (Fig. 2C; Mdn = 64.3%, $p < .001$, $z = 10.69$ for blocks 0 versus 1, Wilcoxon sign rank). The success rate further decreased in block 2 (Fig. 2C; Mdn = 59.0%, $p < .001$, $z = 5.74$ for blocks 1 versus 2, Wilcoxon sign rank) and even further in block 3 (Fig. 2C; Mdn = 56.9%, $p = .015$, $z = 2.44$ for blocks 2 versus 3, Wilcoxon sign rank). Therefore, our progressive series of punishments increasingly stressed the participants' cognitive demands for avoiding the competitor.

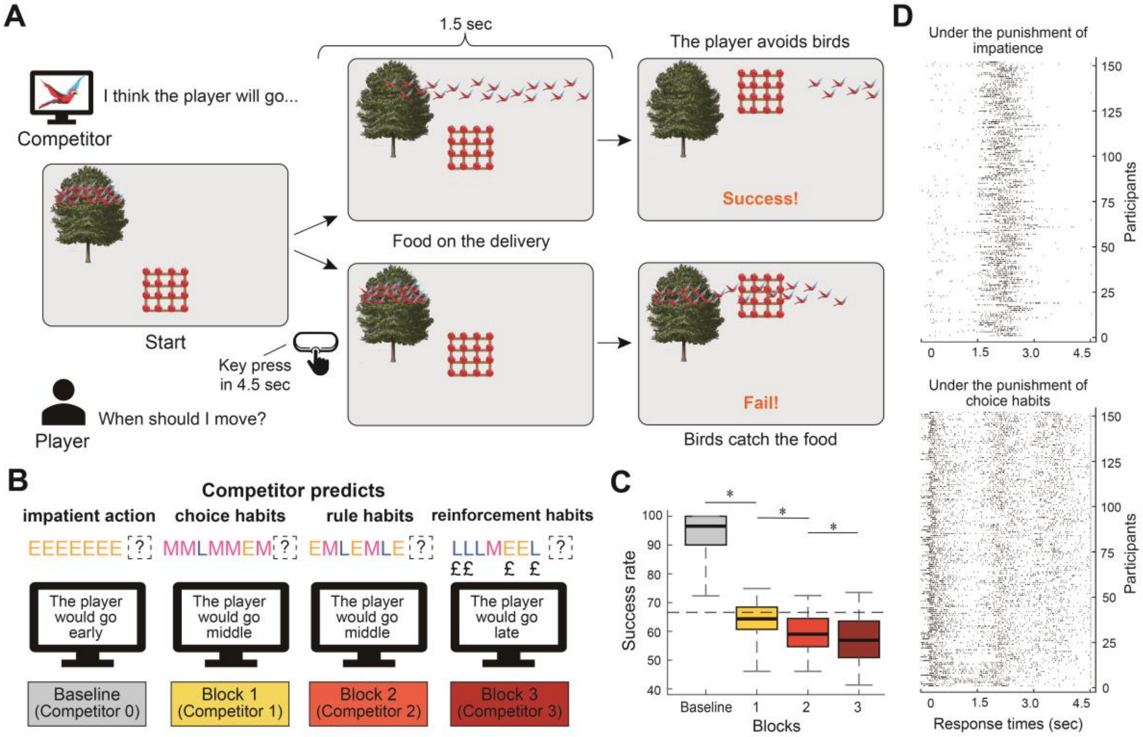


Figure 2. Virtual competitive environments for penalising habitual actions. **A**. A trial sequence. A participant decides when to throw food within a 4.5 second time window. The food is delivered at the top-centre of the screen 1.5 sec after a key press. A virtual competitor (i.e., a flock of birds) attempts to intercept the food by adjusting the time at which it flies out of a tree. Participants win a trial if they avoid being caught by the birds. **B**. Experimental (game) design. The virtual competitor predicted which time interval participants would initiate the delivery based on their past behaviour. On each trial, the competitor punished one of three intervals, 1) early throw (0–1.5 sec), middle throw (1.5–3.0 sec) or late throw (3.0–4.5 sec). An example sequence of action intervals is shown. In the baseline block, the early interval, associated with impatience, was punished. In block 1, *standard choice habits* that favoured one interval over all others were punished (e.g., the middle interval). In block 2, *transition habits* (i.e., sequential pattern) were punished. For example, if the participant went early, middle, late, early, middle and late, the last early action would likely prime the next middle interval. In block 3, *reinforcement habits* (i.e., outcome dependence) were punished. In the example, the repetition of the same interval likely follows from a reward while a change in interval likely follows from no reward. The rewarded last late action would prime the next late interval. **C**. Success rate of avoiding the birds against each class of competitor. A dashed line denotes the chance level. For each box, the central mark represents the median, the edges of the box are the 25th and

75th percentiles and the whiskers are the 2.5th and 97.5th percentiles. * p < significant level after Bonferroni correction, Wilcoxon signed rank. N = 152. **D**. Response times before the punishment of choice habits in the baseline block (upper panel) and during the punishment in block 1 (lower panel). Each small dot represents a reaction time in each trial. The response time data for each participant are aligned in each column.

Do people adapt to punishments of habitual actions?

We next examined the extent to which people could adapt to punishment of different habits. We therefore developed a measure that reflects an individual's tendency towards a trait habit in each habit family. We measured the statistical distance (Kullback-Leibler [K-L] divergence) between the observed probabilities of selecting the early, middle and late intervals, and the probabilities that a habit-free agent would exhibit (right arrow in Figure 3A-C). A lower statistical distance means that the individual is close to the habit-free agent in terms of his choice profile. We call this quantity a decision bias. A decision bias score *before* the punishment of a specific habit reflects an individual's trait habit regarding when to act. We then looked at the *change* in bias score when a given habitual behaviour was punished. We quantified the decision bias score for standard choice habits, transition habits and reinforcement habits, respectively (Figure 3A-C). A greater change in bias score would indicate stronger *adaptive autonomy*, or ability to modulate the trait habit. See *Data analysis* for detailsData analysis.

Looking at response times, the participants began to distribute action times appropriately when Competitor 1 started punishing choice habits (Fig. 2D). Accordingly, their choice bias**—**a statistical distance of the observed choice probabilities from probabilities 0.33 (Fig. 3A)**—**reduced after punishment (a dashed rectangle in Fig. 3D; Mdn = 1.37 for the punishment of impatience versus Mdn = 0.06 for the punishment of choice habits, p < .001, z = 10.69, Wilcoxon sign rank). Competitor 1 did not seek transition patterns from one action to the next and allowed participants to still use *transition habits*. We quantified the sequential bias by considering the extent to which probabilities given the previous action are explained by one's choice probabilities (Fig. 3B). We found that the sequential bias decreased after the competitor pressurised transition habits (a dashed rectangle in Fig. 3E; Mdn = 0.13 for the punishment of choice habits versus Mdn = 0.09 for the punishment of transition habits, p < .001, z = 3.82, Wilcoxon sign rank). Against Competitor 3, the participants were asked to act even more freely to avoid *reinforcement habits*. We evaluated the reinforcement bias by considering the extent to which probabilities given the previous action and the previous outcome are explained by probabilities given the previous action solely (Fig. 3C). The reinforcement bias did not show a significant improvement (a dashed rectangle in Fig. 3F; Mdn = 0.11 for the punishment of transition habits versus Mdn = 0.11 for the punishment of reinforcement habits, *p* = .79, *z* = 0.27, Wilcoxon sign rank).

We further tested the possibility that the participants adapted differently to the influence of positive and negative reinforcements since the neural process after a positive outcome stimulus is different from that after a negative outcome stimulus (Gehring & Willoughby, 2002; Hajcak et al., 2006; Vickery et al., 2011), leading to a stereotypical win-stay lose-shift strategy (Wang et al., 2014). We quantified the positive reinforcement bias and the negative reinforcement bias separately (Suppl. Fig. 1). Nevertheless, we did not find significant improvements in the positive reinforcement bias (a dashed rectangle in Fig. 3G; Mdn = 0.29 for the punishment of transition habits versus Mdn = 0.34 for the punishment of reinforcement habits, *p* = .21, *z* = -1.26, Wilcoxon sign rank) nor in the negative reinforcement bias (a dashed rectangle in Fig. 3G; Mdn = 0.43 for the punishment of transition habits versus Mdn = 0.41 for the punishment of reinforcement habits, *p* = .19, *z* = 1.30, Wilcoxon sign rank). These results suggest that people are able to become more autonomous from standard habitual choices and habitual action transitions but cannot break away from outcome dependencies. That is, people display habits of being guided by reinforced, such as win-stay lose-shift, even when they are discouraged from doing so.

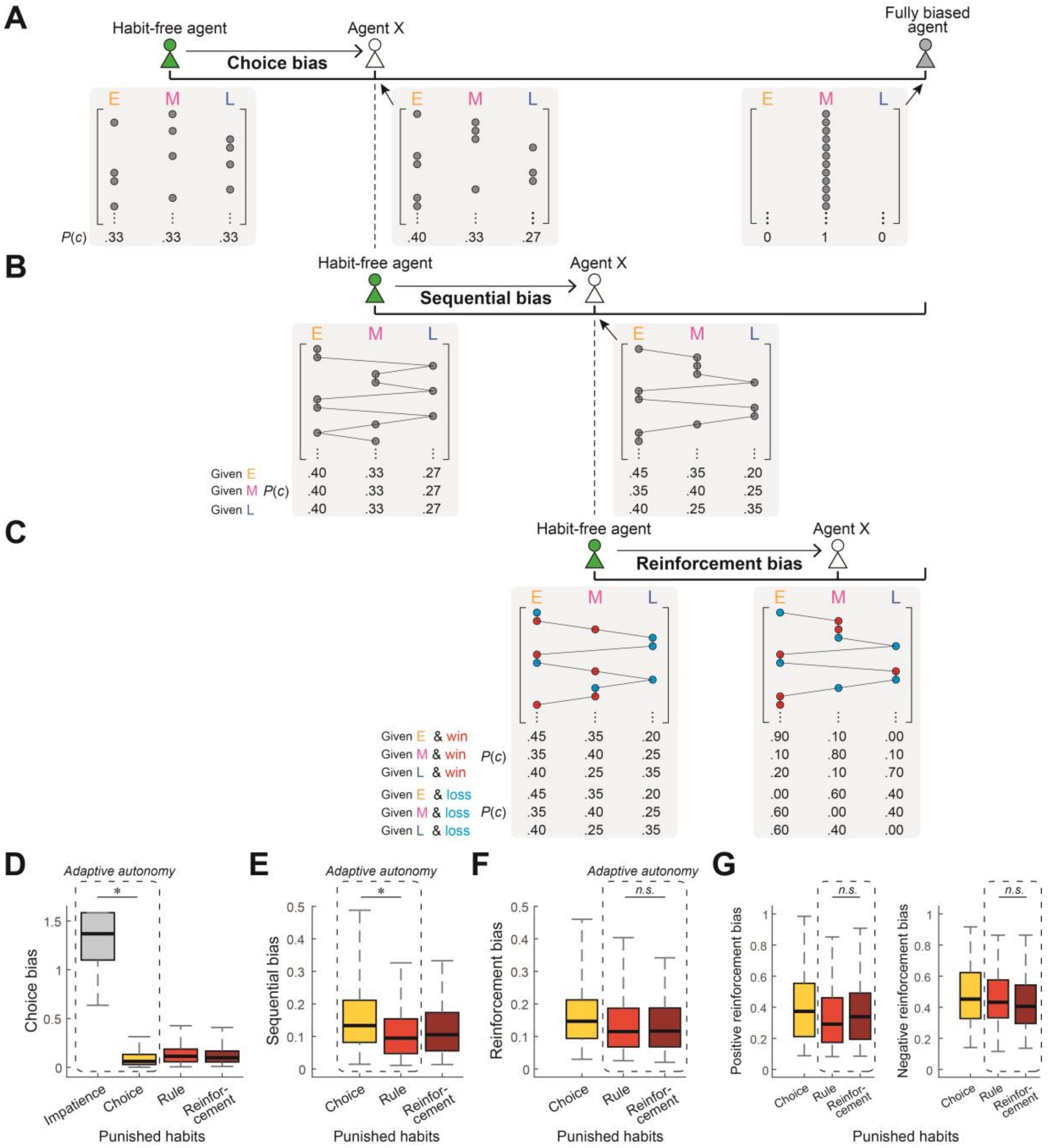


Figure 3. Measuring trait habits and adoptive autonomies in each habit family. In the raster plot, potential actions for the early, middle and late intervals are shown. The numerical values underneath the plot are the probabilities for choosing these three actions. A right arrow illustrates the statistical distance (i.e., pattern similarity) between a habit-free agent and a hypothetical agent X given their choice probabilities. This distance (or decision bias score) is a

proxy for an individual's trait habit. The lower the distance, the smaller the bias. **A.** Choice bias (a proxy for *standard choice habits*). A habit-free agent would select each interval equally often while a fully biased agent would select one interval on every trial. The profile of a participant's choices should be somewhere in-between. **B.** Sequential bias (a proxy for *transition habits*). A habit-free agent is underneath agent X in the panel (**A**) to have the identical choice probabilities to him. The choice probabilities might be larger than the random probabilities 0.33. However, the agent who is free from transition habits would choose their next interval independently of their previous interval: whether the previous interval was early, middle or late does not affect the probabilities of the next interval. Any deviations from such independent choice patterns are considered residual transitions from which participants cannot break (depicted by a right arrow). **C.** Reinforcement bias (a proxy of *reinforcement habits*). A habit-free agent is underneath agent X in the panel (**B**) to have the identical conditional probabilities to him. The agent who is free of reinforcement habits would choose their next interval independently of the previous outcome: whether the previous outcome was success or failure does not affect the probabilities of the next interval. Any such independent choice patterns are considered residual outcome dependencies that participants cannot break (depicted by a right arrow). **D-G.** A dashed rectangle highlights adaptive autonomy as the theoretical difference between a pre-punishment and a post-punishment. * p < significant level after Bonferroni correction, Wilcoxon signed rank. On each box, the central mark represents the median, the edges of the box are the 25th and 75th percentiles and the whiskers are the 2.5th and 97.5th percentiles.

Is there a common factor underlying adaptive autonomies?

A change in decision bias scores between the pre-punishment and the post-punishment phase provides a measure of adaptive autonomy for each habit family (dashed rectangle areas in Fig. 3D-G). We considered the domain-general mechanism of cognitive control, which proposes that proactive, strategic cognitive control shares its control mode across tasks that recruit different cognitive elements (Braver, 2012; Braver et al., 2007; Tang et al., 2022). If adaptive ability to become free of a specific habit (e.g., standard choice habits) generalises to adaptive ability to become free of another habit (transition habits or reinforcement habits), for example because both depend on a common, domain-general mechanism, then we would find a correlation between measures of adaptive autonomy elicited by different types of punishments. Looking at the correlation structure, we did not find strong or even moderate correlations among them, in our sample of 152 participants (Figure 4A). This suggests that the ability to voluntarily regulate one habit is not associated with the ability to regulate another habit. This also suggests that our measurements are separable and evaluate three distinct forms of autonomy conceptualised above, rather than a single common form of autonomy. We also checked the correlation structure between the adaptive autonomy of positive reinforcement bias and that of negative reinforcement bias. There was no strong correlation (Figure 4B). This suggests that the ability to adapt away from a win-stay type behaviour is not associated with the ability to adapt away from a lose-shift type behaviour across participants. To summarise, people seem to recruit distinct cognitive capacities for autonomy when unlearning standard choice habits, transition habits and reinforcement habits.

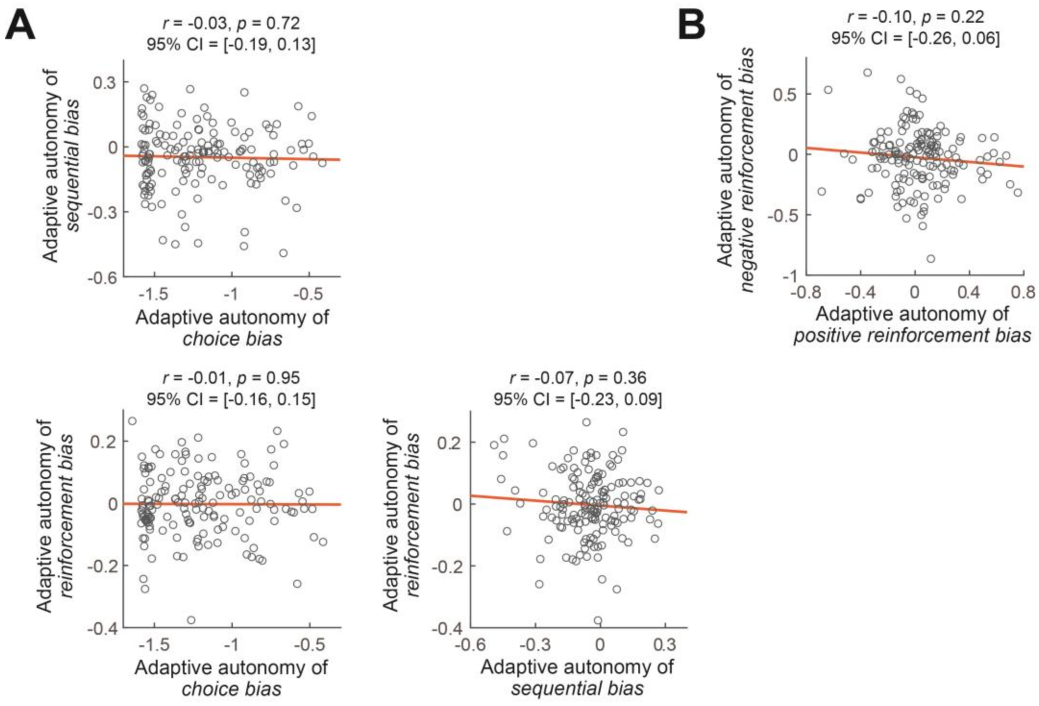


Figure 4. Correlation structure underlying measures of adaptive autonomy. **A.** A measure of adaptive autonomy for each habit family (choice, transition and reinforcement) is quantified as a change in decision bias scores between the pre-punishment and the post-punishment phase (dashed rectangle areas in Fig. 3D-F). In a sample of 152 participants, there is no strong correlation among three measure of adaptive autonomy, suggesting an adaptation is specific to a particular habit family. **B.** A measure of adaptive autonomy for a positive reinforcement bias

is plotted against that for a negative reinforcement bias, quantified as a change in decision bias scores (dashed rectangle areas in Fig. 3G).

A learning process that accounts for behavioural autonomy

What mechanisms could explain how participants adapted to these pressures? How did they learn actions that successfully avoided their competitor? In competitive games, two strategies can be taken to sustain performance. One strategy is stochastic selection by tossing a coin. This mixed strategy helps with unpredictability but does so without interacting with the environment or competitor. An alternative strategy attempts to predict the opponent's next action based on a history of their prior actions (Hampton et al., 2008; Zhu et al., 2012). This strategy is called belief learning (Camerer, 2003) – broadly speaking, the strategy adopted by the virtual competitor is considered belief learning. Belief learners employ an element of mentalizing because they engage in a representation of the actions and intentions of their opponent (Amodio & Frith, 2006; Hampton et al., 2008). Because our task is designed to produce stimulus independence, a BL strategy is a predictive way of achieving stimulus independence (i.e., predict when the birds would fly and avoid them). Thus, we can reason that, as a learning pathway, the participants might learn 1) a feedback-independent stochastic selection strategy, or 2) a feedback-based belief leaning strategy.

We simulated play to test whether the BL strategy is effective at avoiding competitor's predictions. We calculated the reward obtained from simulated choices that BL agents made (see *Simulated play*). We also computed the simulated success rate of a simpler strategy, reinforcement learning, which selects the action that was the most rewarded (Sutton & Barto, 2018). In principle, an RL agent knows whether their choice was rewarded or not and repeats the most rewarded action until it is punished. In contrast, a BL agent knows which option a competitor chose and which options they did not. For instance, if the birds intercepted the early throw, a BL agent reduces the value of the early throw. At the same time, a BL agent can also increase the values of the middle and late throws because the birds did not choose to intercept. An RL agent can only update the value of the throw based on whether it was successful or not. This difference in the internal processes allows a BL agent to update the values of the options quickly, thereby prompting a frequent update of the best option. In the simulated play, both RL and BL strategies achieved sufficient success rates under the punishment of impatience (Fig. 5A; Mdn = 96.7% for RL versus Mdn = 98.3% for BL). However, the BL strategy sustained a higher chance of winning than the RL strategy under the punishment of choice habits (Fig. 5A; Mdn = 60% for RL versus Mdn = 66.6% for BL) and under the punishment of transition habits (Fig. 5A; Mdn = 51.6% for RL versus Mdn = 63.3% for BL). The use of belief learning might account for participants' adaptions to explore new actions.

To address which of the above strategies (stochastic selection, RL and BL) captured our participants' behaviour, we fitted each model to their choice data (see *Computational models*). We also fitted a hybrid learning rule (Camerer & Ho, 1999; Hampton et al., 2008; Zhu et al., 2012) which combines reinforcement learning and belief learning. The stochastic selection model had parameters that captured a participant's choice preference and uncertainty, while the RL and BL models added a learning rate from feedback information. In this nested structure, a better model index relative to the stochastic model would indicate the presence of a feedback process. Across four blocks, we found a clear transition in the best-fitting model (Fig. 5B). According to the summed AICc across participants, the RL model outperformed the stochastic model and BL model under the punishment of impatience. However, the BL outperformed when the competitor started punishing choice habits. The BL model still outperformed when the competitor punished transition habits and reinforcement habits (Fig. 5B). For each model we calculated the number of participants best fit by the model (Fig. 5C) and the protected exceedance probability from the group-level Bayesian model selection (Rigoux et al., 2014; Stephan et al., 2009), which is an omnibus measure of the probability that the model is the best model among tested models. The protected exceedance probability for the RL model to outperform the stochastic model and BL model was close to 100% under the punishment of impatience. The protected exceedance probability for the BL model was close to 100% under the punishment of choice habits and transition habits, and this was 97% under reinforcement habits (Fig. 5C).

As such, the hybrid learning rule fitted the data relatively well in all blocks (Fig. 5B). We recovered the estimates of a relative contribution of belief learning over reinforcement learning from the model fit for the hybrid rule (see *Computational models*). We first checked the robustness of our estimates of the hybrid model by correlating the estimated relative weight parameter to the difference in the AICcs between the RL only model and BL only model. We found a strong positive correlation: the larger the weight placed on BL in the hybrid model, the better the BL only model is (Fig. 5D). We then checked the estimates of the weight parameters across blocks. The weight increased from punishment of impatience to punishment of choice habits (Fig. 5E; Mdn = 0.00 for the punishment of impatience versus Mdn = 0.63 for the punishment of choice habits, p < .001, z = 8.32, Wilcoxon sign rank). That is, belief learning made an important contribution to participants' choices when they were punished for choice habits (Fig. 5E; Mdn = 0.63), transition habits (Mdn = 0.81; versus choice habits, p = .10, z = 1.65) and reinforcement habits (Mdn = 0.74; versus transition habits, p = .96, z = 0.04), respectively. These findings suggest a shift in learning strategies that followed the demands of the competition: participants first used reward-guided behaviour when it sustained the success rate. Then, once the competitor started predicting habit patterns, participants switched to learning successful actions from the opponent's prior actions.

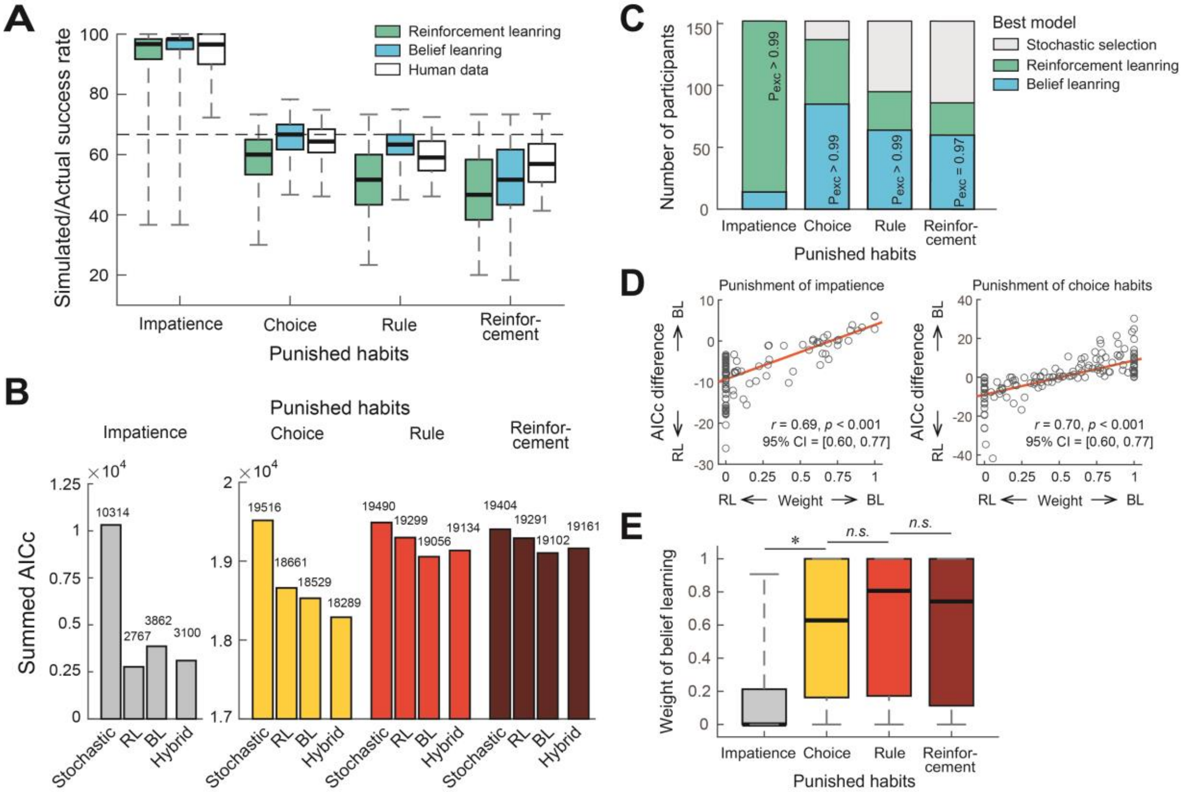


Figure 5. A shift in the strategy to belief learning as competitive demand increases. **A**. Real success rate (white bars) in the actual experiment and fictive success rate (green or blue bars) in simulated play. Agents using the reinforcement learning strategy and agents using the belief learning strategy competed against each class of competitor. **B.** Summed AICc across participants. Lower values of AICc are better. Stochastic: stochastic selection model. RL: reinforcement learning model. BL: belief learning model. Hybrid: hybrid learning rule. **C.** The number of participants best fit by the model. The P_{exc} inserted in the bar denotes the protected

exceedance probability that supports the corresponding model. **D.** Relative weight placed on belief learning over reinforcement learning captured by the hybrid model versus the AICc difference between the RL model and BL model. As the weight parameter increases, the model fit of belief learning improves relative to that of reinforcement learning. **E.** A transition in the relative weight placed on belief learning. * *p* < significant level after Bonferroni correction, Wilcoxon signed rank. **A&E.** On each box, the central mark represents the median, the edges of the box are the 25th and 75th percentiles and the whiskers are the 2.5th and 97.5th percentiles.

A shift in learning strategies enhances behavioural autonomy

Because belief learning is a faster learning process than reinforcement learning, our simulated play shows that belief learning indeed induces a smaller choice bias than simple reinforcement learning (Suppl. Fig. 2), leading to better performance (Fig. 5A). We therefore examined whether the shift in learning processes accounts for achieving a smaller choice bias in the participants' data. To this end, we used a bivariate latent change score (LCS) model (Carpenter et al., 2019; Kievit et al., 2018; Kievit et al., 2017; McArdle, 2009). LCS models conceptualize the change in score between one time point (before punishment) and the next time point (under punishment) as a latent change factor (see *Latent change score model*). Under a bivariate LCS model, two factors influence the change score. The first is the extent to which the reduction in bias is explained by the initial bias before punishment, which is termed auto-regression in LCS models. We might expect to see a negative auto-regressive effect as a consequence of a scale attenuation: individuals who started off with a larger bias score could potentially have a greater reduction in the bias. While individuals those who started off with a smaller bias score could have a smaller reduction in the bias because of the lower limit of the scale. The second is the extent to which the reduction in bias is explained by the initial weight on belief learning, which is termed cross-coupling. A bivariate LCS model would reveal a negative cross-coupling effect if individuals who attempted to mind-read the competitor's strategy (i.e., having a large BL weight before punishment) gained a greater reduction in the bias. Moreover, by having two latent change factors, a bivariate LCS model estimates correlated change: the degree to which the reduction in bias co-occurs with the change in weight (i.e., a shift in learning strategies). If participants adapted their behaviour to gain a smaller bias score by simply behaving randomly, the reduction in bias would not co-occur with the change in weight. If, on the other hand, participants achieved a smaller bias score by initiating the prediction of the competitor's strategy, these two changes would co-occur, and then the bivariate LCS model would reveal a negative correlated change: gaining a greater weight is associated with gaining a smaller bias.

We investigated this inter-relationship by fitting the bivariate LCS model. Our observed variables were the choice bias score and the belief learning weight estimated under the hybrid learning rule. To estimate the latent change factors, we used these scores on the punishment of impatience in the baseline and the punishment of standard choice habits in block 1, because the participants were asked to change their choice habits between these two conditions. Figure 6A illustrates fitted paths (significant paths are shown as thicker lines) from the pre-punishment phase (T0: punishment of impatience) to the post-punishment phase (T1: punishment of standard choice habits). This model, with fixed intercepts, shows a fit close to the saturated (i.e., perfect) model ($\chi^2(4) = 0.00$; RMSEA < 0.001, 90% confidence interval = [0.000, 0.000]; CFI = 1.00; SRMR = 0.00; AIC = -325.1). We found auto-regressive effects in both the choice bias score and BL weight score possibly because of a scale attenuation: greater reductions in the choice bias were found in individuals who started off with a large bias, and greater gains in the BL weight were found in individuals who started off with a low weight. There were no significant cross-coupling effects: greater reductions in the choice bias were not associated with individuals who attempted to mind-read the competitor's strategy before punishment.

However, we found a medium negative correlated change between the change in the choice bias and the change in the BL weight. This suggests that individuals who gained a greater shift from reinforcement learning to belief learning exhibited a greater adaptation in choice habits (Fig. 6B; standardised coefficient = -0.347, 95% confidence interval = [-0.480, -0.199]). When we removed the cross-coupling paths from the model in Figure 6A, the model still fitted the data well ($\chi^2(6) = 2.37$; RMSEA < 0.001, 90% CI = [0.000, 0.053]; CFI = 1.00; SRMR = 0.03; AIC = -326.8). However, removing the cross-coupling paths as well as the path for the correlated change produced a bad model fit ($\chi^2(7) = 23.80$; RMSEA = 0.126, 90% CI = [0.071, 0.185]; CFI = 0.68; SRMR = 0.10; AIC = -309.0). To summarise, the computational model suggests that people start forming their beliefs about when the competitor is going to act and when they should act, after the competitor start responding to their own actions and punishing their own choice habits. Together with structural equation modelling, the results suggest that forming beliefs about the competitor's strategy helps to enhance adaptive autonomy in avoiding choice habits.

We validated this finding in the following analysis. First, we used the choice bias score and the difference in AICcs between the RL only model and BL only model as our observed variables. We used these scores on the punishment of impatience in the baseline and the punishment of standard choice habits in block 1. We still found that the reduction in the choice bias co-occurs with a shift in strategies toward belief learning (Suppl. Fig. 3A). This result validates that the correlated change in Figure 6 is unlikely to be due to the boundaries of the parameter estimates under the hybrid learning rule.

Second, to validate whether the competitor's action indeed affected the participant's action, we randomised the competitor's action in the sample we fitted the hybrid rule. This permutation operation indeed disrupted the estimates of the BL weight (Suppl. Fig. 4A). In the permutation sample, we did not find evidence that the reduction in the choice bias co-occurs with a shift in strategies toward belief learning (Suppl. Fig. 4B&C). This result validates that the competitor's action likely affected the participant's action, and that the shifting to the BL strategy was likely associated with changing a pattern of choice.

Lastly, we checked the possibility that the reduction in the sequential bias or the reinforcement bias co-occurs a shift in strategies. We did not find evidence that the reduction in the sequential bias co-occurs with a shift in strategies toward belief learning, when the transition habits were punished in block 2 (Suppl. Fig. 3B). Nor did we find that the reduction in the reinforcement bias co-occurs with a shift in strategies toward belief learning, when the reinforcement habits were punished in block 3 (Suppl. Fig. 3C). These follow-up analyses confirm that our finding is not simply explained by the fact that the participant's data was used to estimate both a measure of decision bias and parameters of the hybrid model.

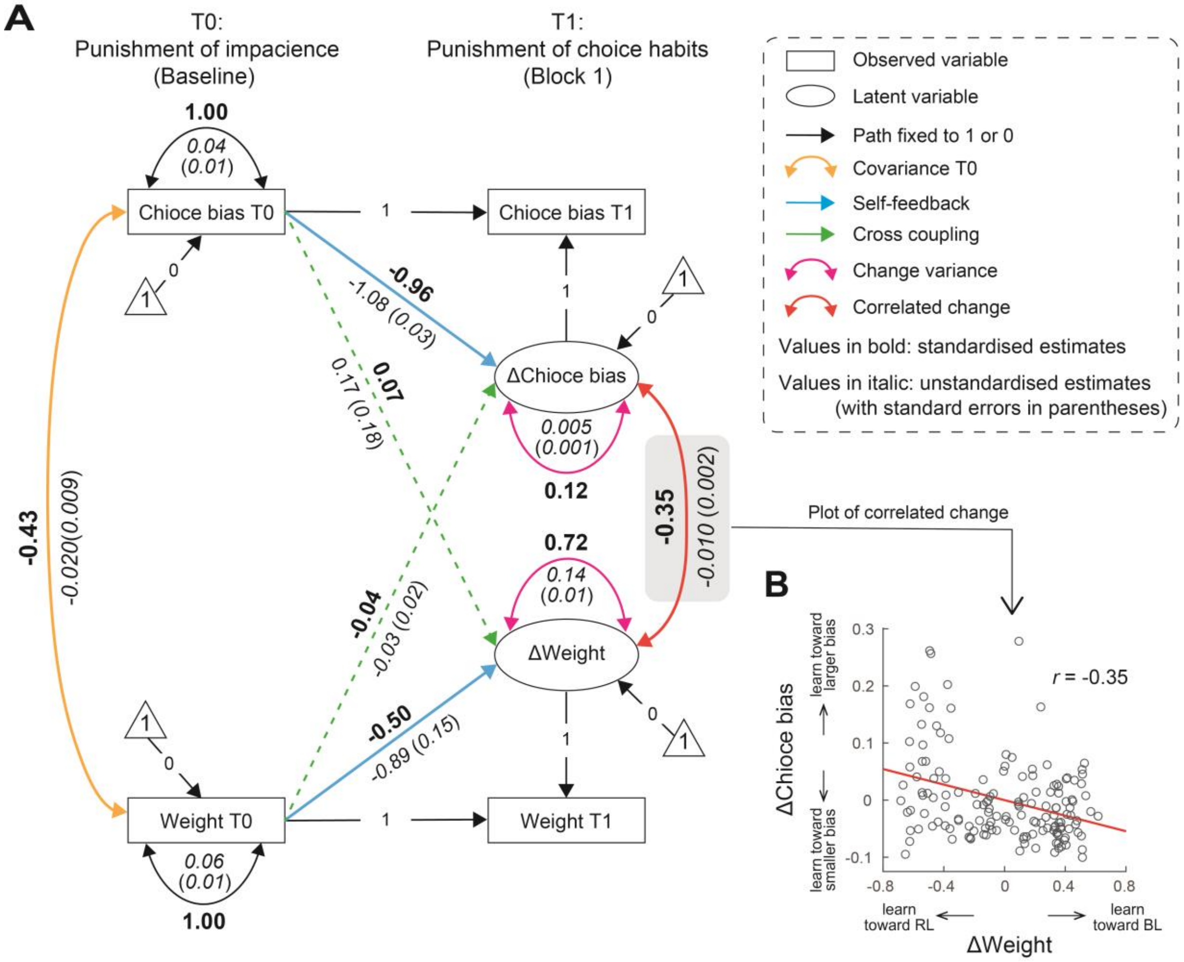


Figure 6. The bivariate latent change score model of adaptive autonomy and strategic learning. **A**. Estimated parameters. The change score in the choice bias was modelled as a latent factor between the score before the punishment of choice habits and under the punishment. The change score in the belief learning weight recovered from the hybrid leaning rule was modelled similarly. Values in bold are standardised parameter estimates, and values in italic are un-standardised parameter estimates (with standard errors in parentheses). Solid lines indicate that the parameter is significant at $p < .05$. **B**. Scatter plot of correlated change. A greater adaptation in avoiding standard choice habits is correlated with a greater shift in the learning process from reinforcement learning to belief learning.

Discussion

The human capacity for autonomous behaviour is widely asserted, and is fundamental to many modern societies, but its cognitive basis is not well understood. We developed an experimental paradigm that implicitly elicits autonomous behaviour, and we measured the extent to which people could express autonomy by adapting their behaviour to free themselves from different types of habits. We found that people can become autonomous of standard choice habits and transition habits, but are limited in their ability to become free of reinforcement habits. We further showed, in a large sample, that adaptive autonomy with respect to choice habits is uncorrelated with adaptive autonomy with respect to transition habits and that to reinforcement habits. This suggests distinct cognitive modules for these three forms of autonomy, rather than a common module or a single form. We further demonstrated the link between adaptive autonomy and understanding the state of the environment: building beliefs about the environmental, in our case, about a competitor's strategy, can enhance adaptive autonomy.

Measures of autonomy

Traditional experimental psychology struggles to investigate autonomy because traditional experiments in the studies of volition involve instructing people what they should do (Baddeley, 1966; Brass & Haggard, 2007; Fleming et al., 2009; Jahanshahi et al., 1995; Libet et al., 1983). The few studies that have examined human autonomous behaviour typically involve competitive contexts (Forder & Dyson, 2016; Wang et al., 2014; Wong et al., 2021), and have not considered subtypes of autonomy. We conceptualized three forms of autonomy, as freedom from three cognitively distinct types of habit: *standard choice habits*, *transition habits* and *reinforcement habits*. We attempted to evoke autonomous behaviour of each of these three kinds using a common game-like context, and varying the competitor's strategy to punish a lack of each type of autonomy. Using a statistical distance measure derived from information geometry, we developed quantities that approximate a person's trait for each habit. We quantified the extent to which people could break a specific habit when punished. A covariance matrix underlying three adaptive autonomy measures showed no strong correlations. This supported the idea that people express three distinct forms of behavioural autonomy. In tasks where participants react to external stimuli quickly, it has been suggested that a domain-general top-down control is used to solve different cognitive tasks (Braver, 2012; Braver et al., 2007; Tang et al., 2022). However, in free, stimulus-independent action, our results suggest that a domain-specific top-down control independently regulates each particular form of autonomy: there are multiple ways to act freely, and it is important to consider *from what* an agent is free. We studied choice biases, sequential biases and reinforcement biases, but other biases to free action doubtless also exist. We showed that, for example, an agent who becomes increasingly free from choice bias may be unable to free themselves from the biasing effects of reinforcement.

Relevance to classical neuropsychological tasks

Our task takes a neuropsychological perspective on volitional behaviour and evokes phenomena that neuropsychologists have traditionally studied using arbitrary, open-choice tasks. For instance, our measure of choice bias is related to the capacity to inhibit a prepotent, impulsive action (Mischel et al., 1972). People usually place costs on waiting, preferring earlier rewards; a form of temporal discounting (Story et al., 2014). The sequential bias we measured reflects executive control and working memory which are assessed using random number generation tasks (Baddeley et al., 1998; Jahanshahi et al., 2000). In these tasks, people cannot sufficiently randomise numbers and tend to seek simple rules such as repeating a digit (e.g., 1,1,1), counting a digit in a natural sequence (e.g., 1,2,3) or larger with smaller inter-digit gaps (e.g., 1,9,2) (Baddeley, 1966; Bar-Hillel & Wagenaar, 1991). In contrast, the capacity to avoid the reinforcement bias is related to voluntary regulation of reward-seeking behaviour (Bechara et al., 1994; Lejuez et al., 2002) and a balance between exploitation and exploration (Cohen et al., 2007). These classical preferences are typically robust after repeating the task (Neiman & Loewenstein, 2011; Ota et al., 2016) or after explicitly knowing one's own behavioural trait (Ota et al., 2019). We used a competitive game to set environmental constraints so that people should avoid such preferences in order to get rewards. Our results demonstrate that people can balance choice frequencies and break transitions between actions. Although we cannot directly compare our results with those of neuropsychological studies, our results do suggest that, when pressurised, people can behave randomly and autonomously more than suggested by the traditional neuropsychological literature.

In contrast, we found people could not avoid reinforcement habits. Neither positive reinforcement bias nor negative reinforcement bias changed when penalised. A stereotypical win-stay lose-shift behaviour has been shown in competitive games (Ota et al., 2020; Wang et al., 2014). In particular, people are less flexible in changing lose-shift behaviour than win-stay behaviour when adopting to new game rules (Forder & Dyson, 2016; Sundvall & Dyson, 2022). The experience of a negative outcome generates a quicker decision and an impulsive response on the next trial (Dyson et al., 2018). Indeed, event-related brain potentials show a greater negative amplitude after negative outcome feedback than after positive outcome feedback (Gehring & Willoughby, 2002; Hajcak et al., 2006). We note that individuals vary in their ability to avoid habitual behaviour after negative reinforcement. These individual differences are associated with post-error reaction times. Individuals who make quicker decisions after a loss than after a win show a poorer performance than individuals who make slower responses after a loss (Dyson, 2021). Therefore, overcoming impulsivity after a loss may be a key aspect of volitional control for humans.

Sustaining autonomy during interactions with the world

Our model comparisons showed that participants did not achieve adaptive autonomy simply by behaving randomly and stochastically. Rather, their strategy depended on the reward assigned and the choices of the competitor. Reinforcement learning updates the best action by a reward prediction error while the belief learning updates the best action by an action prediction error, defined as a difference between the expected action competitor would take and the actual action taken. Both reinforcement learning and belief learning can contribute to volitional, self-generated actions, because both are determined by internal representations of expected values, rather by an immediate stimulus (Frith, 2013). Furthermore, belief learning is considered to recruit a mentalizing or an implicit understanding of what the other agent would do (Amodio & Frith, 2006; Hampton et al., 2008). These neural substrates are often found in separate neural networks: the reward prediction errors are encoded in the ventral striatum (McClure et al., 2003; O'Doherty et al., 2004; Zhu et al., 2012) while the prediction errors about the state of the environment are encoded in several areas including the rostral anterior cingulate, the medial prefrontal cortex and the posterior superior temporal sulcus (Hampton et al., 2008; Zhu et al., 2012).

Mentalising is a key cognitive component recruited in competitive games (Hampton et al., 2008; Zhu et al., 2012). We found that people switch their strategy from reinforcement learning to belief learning when the competitor started predicting their upcoming action. This result suggests that people implicitly learned the likelihoods of the competitor's actions based on a sampling of their past actions. Critically, belief learning was associated with enhanced autonomy. In particular, successful adaptation in avoiding standard choice habits was associated with a shift in learning: those individuals who shifted to learn from the likelihoods of the competitor's actions rather than from reinforcement alone were able to gain greater adaptive autonomy. We cannot tell whether participants discovered and explicitly represented the punishment rules used by the competitor, but we speculate that explicitly understanding the constraints on behaviour might be associated with increased autonomy.

To conclude, we have developed a new experimental paradigm and analysis pipeline to study when and how human actions can become autonomous. We propose a new theoretical construct of *adaptive autonomy*, meaning the capacity to free one's behavioural choices from constraints of habitual responding, when a particular habit becomes dysfunctional, for example due to environmental changes like the competitive pressure in our game scenarios. We have shown that people can indeed express adaptive autonomy, and that they do so by reducing habits of choice, habits of rule-based sequential action and habits of being guided by reinforcement. These appear to be three distinct forms of adaptive autonomy, rather than a single common strategy such as randomness. We show that becoming free from the effects of reinforcement is particularly difficult. Finally, by showing that belief learning plays an important part in boosting autonomy, we show a strong connection between autonomous action and mentalising abilities.

Limitations

The three habit forms we tested were hierarchical. The standard choice habit – favouring one action over all others – is more general, while the reinforcement habit is more specialised. Therefore, we ordered the blocks so that the competitive game algorithms could penalise habits progressively and serially. Thus, each block implicitly required participants to act more freely and unpredictably than the preceding block. This fixed order may limit the generalisability of our results, but the order we used is the most reasonable. The differences we observed between the different forms of adaptive autonomy could be confirmed in further between-participant studies.

Empiricist view versus nativist views of human autonomy

Our work is broadly compatible with an empiricist view of "free will" as opposed to a nativist view. In our view, some of the key attributes historically associated with "free will", such as the ability to act endogenously and purposefully, can be acquired, or at least adapted, through experience. This adaptation requires people to make novel, non-habitual, 'smart' actions in certain situations. We found that people were more or less successful in adapting their trait habits both at the individual level and at the level of different punishments. An individual's degree of autonomy is unique and contingent on environmental constraints. A strongly nativist view would suggest that autonomy is a state that occurs inside an individual's mind and is independent of the external world. However, our results imply that being sensitive to the contingencies of the external environment, and its restrictions on one's own actions is key to autonomy. In this sense, autonomy can be seen a reasoned, goal-oriented response that occurs within an environmental context.

Methods

Participants

One hundred and fifty-nine participants (age range = 18–45, M = 29.5 yo, SD = 7.2) were recruited online via the Prolific website (https://www.prolific.co/). Participants received a basic payment of £3.75 for their participation in a 30 minute experiment. They earned a bonus of up to £4 based on their performance on the task. There were 95 female participants and 64 male participants. Recruitment was restricted to the United Kingdom. Seven participants were excluded from the analysis (Suppl. Info.) and the remaining 152 participants were analysed. All procedures were approved by the Research Ethics Committee of University College London. Participants gave informed consent by checking and validating the consent form.

Experimental design

Apparatus. We used the JavaScript library jsPsych (de Leeuw, 2015) and the plugin jsPsych-psychophysics (Kuroki, 2021) to program the task and hosted the experiment on the online research platform Gorilla (https://gorilla.sc/) (Anwyl-Irvine et al., 2020), which participants could access through their browser on their own computer. We assumed that monitor sampling rates were typically around 60 Hz, with little variation across computers (Anwyl-Irvine et al., 2020). The size and position of stimuli were scaled based on each participant's screen size which was automatically detected. The size of stimuli reported below are for a monitor size of 15.6" (view point size, width x height: 1536 x 746 pixels).

Stimuli and task. Each trial started with a fixation cross, which appeared for 0.6–0.8 seconds. The images of a tree, a flock of birds and a basket containing apples then appeared (Fig. 2A). A tree (width x height: 307 x 375 pixels) was shown on the left of the screen and a flock of birds (width x height: 30 x 22 pixels each) were located on the tree. A rectangular basket of apples (width x height: 153 x 167 pixels) was presented in the bottom centre. After the fixation cross disappeared and all images appeared, the participants were given 4.5 sec to throw the food. Pressing a key initiated delivery of the food to a storage location which was located at 447 pixels forward from the start point. This delivery took 1.5 sec. We programmed the birds to attempt to intercept and catch the food. The birds on each trial were designed to intercept the food thrown within one of three intervals: 1) early throw (0–1.5 sec), 2) middle throw (1.5–3.0 sec) or 3) late throw (3.0–4.5 sec). After their departure, it took approximately 0.25 sec for each bird to reach the storage location; the birds passed through that point. The participants competed with the virtual competitor, aiming to deliver food before or after the birds reached the storage location. We counted whether one of the birds overlapped with the food when the delivery was completed (at the offset of moving). If this was the case, the food was caught, and the participant lost a trial. If not, the food was delivered without it being caught, and the participant won a trial. If no response was submitted before 4.5 sec, the food was launched automatically, and a trial was terminated as a timeout. Finally, we provided a feedback message: “Success!”, “Fail!” or “Timeout!”, which lasted for 1.0 sec. The next trial then started with a fixation cross.

In the instructions, we emphasised the following points. First, merely reacting to the absence of a stimulus – the birds resting in the tree – will not win the game because the birds can travel much faster than the food. Second, merely waiting for the birds to pass is not a solution because of the time constraint. Third, the birds’ flying interval is not the same on every trial, nor is it random. Instead, the birds can learn when the participant is likely to throw the food. Therefore, it is important to predict when the birds will likely fly and to randomise your throw times in order to avoid the competitor’s prediction.

Procedure. Participants first received the instructions and viewed a set of demonstrations about the task. Following some practice trials, the participants completed four blocks of the game with a 1-minute break between blocks. The baseline block lasted 2.5 minutes while the remaining blocks 1, 2 and 3 lasted 5 minutes each. The participants got as many throws of the food as they could in the 2.5 or 5 minutes. The participants could check how much time was left in each block. We used time, and not trial number, to terminate each block so that participants did not respond immediately on every trial, finishing the game early. The bonus payment was determined by the percentage of throws that successfully avoid birds and was paid up to £1 for each block: if 40 out of 60 throws are successful, we paid £1 x 66.6% = £0.66

(average bonus, baseline: £0.94; block 1: £0.63; block 2: £0.59; block 3: £0.57). The success rate and the timeout rate were included in the feedback. Nevertheless, we assumed that some participants might consume time by not focusing on the game. To prevent this, we encouraged participants to sustain the proportion of timeout trials under 5%.

Competitor design

The virtual competitor design was primarily inspired by primate work (Barraclough et al., 2004; Lee et al., 2004) and by rat work (Tervo et al., 2014). We programmed the learning algorithm (i.e., birds) to seek out behavioural patterns in the participant's choice history and to pressure participants into novel behaviour. The participants could decide the time to act between 0 sec (as soon as birds and food appeared) and 4.5 sec (until timeout). To make the competitor's prediction simple, we discretised the time window into three intervals, 1) early interval (0–1.5 sec), middle interval (1.5–3.0 sec) and late interval (3.0–4.5 sec). Given past behaviour, the competitor predicted which response interval a participant was likely to select. Accordingly, the two other intervals were primed for winning: if the participant threw the food during the interval predicted by the competitor, the participant lost. If the participant threw the food during one of two other intervals, the participant won. We adjusted the birds' departure times by taking their travel time (0.25 sec) and the food delivery/travel time (1.5 sec) into account: if a prediction was made on the late interval, the birds departed from the tree at the period of 4.25–5.75 sec, and they reached the delivery point during 4.5–6.0 sec to catch the food when it was delivered.

We designed four distinct competitors (Fig. 2B). First, in the baseline block, Competitor 0 punished participants for being impatient. In this block, the birds blocked the early throw on every trial. Thus, the stimulus-absence behaviour corresponded with waiting until the middle interval. Competitor 0 measured the volitional control to resist immediacy or external triggers (Haggard, 2019). Second, in block 1, Competitor 1 predicted *standard choice habits* (choice preferences) – which interval the participant is going to select –. On each trial, a history of the participant's past ten choices was used to estimate the probabilities of selecting the early, middle and late interval. The choice probabilities were then used to generate the competitor's prediction on the upcoming choice. For instance, if the participant chose the early interval seven times, the middle interval twice, and the late interval once, the competitor penalised the early interval 70% of the time, the middle 20% of the time and the late 10% of the time. Thus, Competitor 1 required participants to balance their general choice frequencies.

In block 2, Competitor 2 sought out *transition habits* (sequential patterns) – which interval the participant is going to select after the participant made a particular response –. A history of the past 60 trials was used to estimate the conditional probabilities of selecting three intervals given the previous reaction time. The estimated probabilities were conditioned on the last reaction time ± 0.5 sec. Suppose a participant took 2.5 sec to act in the previous trial. Competitor 2 might discover that, in the past, the participant chose the early interval twice, the middle interval twice, and the late interval six times after the participant had acted in 2.0-3.0 sec. In this case, Competitor 2 penalised the late interval 60% of the time. We assumed that using the previous response time (i.e., continuous variable) is more powerful to predict the next response than using the previous response interval (i.e., categorical variable). Competitor 2 pressured participants in avoiding habitual transition patterns. Finally, in block 3, Competitor 3 punished *reinforcement habits* (outcome dependence) – which interval the participant is going to select after the participant made a particular response and won a trial or lost a trial –. Competitor 3 used the same search algorithm as Competitor 2 with the exception that they conditioned the search on the last reaction time and the last outcome. Competitor 3 required participants to act independently from the previous outcome.

Data analysis

Because the birds intercepted one of the three response intervals, we mainly analysed the data that was discretised into 1) the early response: responding in 0–1.5 sec, 2) the middle response: responding in 1.5–3.0 sec, 3) the late response: responding in 3.0–4.5 sec (including timeout).

Quantifying trait habits. Statistical distance is a standardised way of measuring the extent to which the observed probability distribution is different from the target probability distribution. We calculated the Kullback-Leibler divergence to quantify the extent to which the participant's choice probability distribution is different from the choice probability distribution that a habit-free agent would exhibit, a proxy of three habit families. See Figure 3.

1) Choice bias. Competitor 1 punished standard choice habits in selecting one interval more often than the other two. The probabilities of choosing the early, middle and late interval for a habit-free agent would be 0.33, respectively. We computed the choice probabilities $P(c)$ given a history of intervals each participant chose in each block. The K-L divergence is then

$$D_{KL\ choice\ bias} = \sum_{c\ \in\ E,M,L} P(c) \log_2 \left(\frac{P(c)}{0.33} \right)$$

2) Sequential bias. Competitor 2 punished transition habits on the top of choice habits. Similar to computing the choice probabilities, we computed the conditional probabilities of choosing the early, middle and late interval given the interval chosen on the previous trial $P(c|c_{-1})$. We measured the K-L divergence of these participant's conditional probabilities from the participant's choice probabilities. The K-L divergence for each previous interval c_{-1} is computed as

$$D_{KL\ c_{-1}} = \sum_{c\ \in\ E,M,L} P(c|c_{-1}) \log_2 \left(\frac{P(c|c_{-1})}{P(c)} \right).$$

The total K-L divergence as a weighted sum is then

$$D_{KL\ sequential\ bias} = \sum_{c_{-1}\ \in\ E,M,L} P(c_{-1}) \cdot D_{KL\ c_{-1}}$$

Since we conditioned the K-L divergence on the previous interval chosen, we took the proportion of observing that situation into account, and we weighed each divergence by this prior probability. The target probabilities (i.e., habit-free agent) were set to be the participant's own choice probabilities, rather than purely stochastic choices 0.33. Therefore, $P(c|c_{-1})$ becomes equivalent to $P(c)$ and the K-L divergence becomes zero, as long as the participant selects three intervals independently from the previous choice (even if the participant favours one interval). By this way, we quantified the deviation of patterns associated with the previous choice from sequential patterns logically expected from the participant's own choice probabilities. Competitor 2 specifically detected and punished this conditional dependence.

3) Reinforcement bias. Competitor 3 punished reinforcement habits on the top of choice habits and transition habits. Similar to computing the choice probabilities, we computed the conditional probabilities of choosing the early, middle and late interval given the interval chosen and the outcome obtained on the previous trial $P(c|c_{-1}, o_{-1})$. We measured the K-L divergence of these participant's conditional probabilities from the participant's conditional probabilities given the previous interval solely. The K-L divergence for each previous interval c_{-1} and each previous outcome o_{-1} is computed as

$$D_{KL\ c_{-1},o_{-1}} = \sum_{c\ \in\ E,M,L} P(c|c_{-1}, o_{-1}) \log_2 \left(\frac{P(c|c_{-1}, o_{-1})}{P(c|c_{-1})} \right).$$

The total K-L divergence as a weighted sum is then

$$D_{KL\,reinforcement} = \sum_{c_{-1} \in E,M,L} \sum_{o_{-1} \in \substack{success \\ fail}} P(c_{-1}, o_{-1}) \cdot D_{KL\ c_{-1},o_{-1}}$$

Since we conditioned the K-L divergence on the previous interval chosen and the previous outcome obtained, we took the proportion of observing that situation into account, and we weighed each divergence by this joint prior probability. The target probabilities (i.e., habit-free agent) were set to be the participant's own conditional probabilities given the previous interval solely. Therefore, $P(c|c_{-1}, o_{-1})$ becomes equivalent to $P(c|c_{-1})$ and the K-L divergence becomes zero, as long as the participant selects three intervals independently from the previous outcome (even if the participant's choice depends on the previous interval). By this way, we quantified the deviation of patterns associated with both the previous choice and the previous outcome from patterns logically expected from the conditional dependence on the previous choice solely. Competitor 3 specifically detected and punished this outcome dependence. We also quantified the positive reinforcement bias and the negative reinforcement bias, separately (Suppl. Fig. 1). We computed the K-L divergence of the conditional probabilities given the previous interval and the previous win only or the previous loss only from purely stochastic choices 0.33. Here the statistical distance can be argued as the distance between the participant's post-win behaviour or post-loss behaviour and the habit-free agent who is purely random. These measures were used to generate Figure 3.

Statistical analysis. We tested the performance difference by Wilcoxon signed rank test. The alpha level of 0.05 was corrected by the number of tests we performed in each class of test (Bonferroni correction).

Computational models

Reinforcement learning. We tested a reinforcement learning (RL) model in which an action value is updated via a Rescorla-Wagner rule (Sutton & Barto, 2018). On each trial, an RL agent selects an action from the early, middle or late interval $a \in E, M, L$. For an action a selected on a trial t, the value of action a is updated by a prediction error δ:

$$\delta_t = r_t - V_t(a)$$

where r_t is the actual reward received (1 for successfully avoiding birds and 0 for failure) and $V_t(a)$ is the current expected reward for that action. The reward prediction error δ_t is then used to update the value of the selected action, weighted by the learning rate α

$$V_{t+1}(a) = V_t(a) + \alpha\delta_t.$$

Belief learning. In a belief learning (BL) model, a BL agent infers the opponent's state of mind – what option the opponent is going to select – and decides on the action that maximises the expected reward (Camerer, 2003; Hampton et al., 2008; Zhu et al., 2012). Actions $a' \in E, M, L$ are available for the competitor to choose. For each action $a \in E, M, L$ on trial t, the value of that action is updated by a prediction error

$$\delta_t = r_t - V_t(a)$$

where $r_t = -1$ if a' is same as a (i.e., the competitor selects the same response interval as the participant) while $r_t = 0$ if a' is different from a (i.e., the competitor selects a different response interval). This prediction error is the difference between the current expected value and the negative reward derived from the competitor's current choice. Therefore, the updated expected value of action reflects the likelihood of the competitor's choice: the larger the value, the less likely the competitor choose. The same rule with the RL model was used to update the value of action a, with the exception that the values of all three intervals were updated on every trial. Suppose that the birds repeatedly selected the early interval to intercept the food. A direct observation of the birds' flight at the early interval decreases the value of the early interval. At the same time, this observation implies that the birds did not fly at the middle nor the late interval. This increases the values of these intervals.

Hybrid learning. We modelled the hybrid learning rule (aka. experience weighed attraction) as a combination of reinforcement learning and belief learning (Camerer & Ho, 1999; Hampton et al., 2008; Zhu et al., 2012). After updating the value of the action in each learning process, the hybrid rule combines the values of the action such that

$$V_{t+1}(a) = (1 - w) \cdot V_{t+1}^{RL}(a) + \ w \cdot V_{t+1}^{BL}(a)$$

where one additional free parameter w is used to weigh the relative contribution placed on belief learning over reinforcement learning.

For all models, the action values were converted into the choice probabilities using the soft-max function to simulate action selection,

$$P_t(a) = \frac{e^{\ \beta \cdot (V_t(a) + b(a))}}{\sum_{a \in E,M,L} \ e^{\beta \cdot (V_t(a) + b(a))}}$$

where $P_t(a)$ is the probability of choosing the interval a. The inverse temperature parameter β scales the relative difference between the choice probabilities, which scales decision uncertainty. We added the decision preference term b with an exponential temporal discounting (Story et al., 2014):

$$b(a) = e^{-\rho \cdot T(a)}$$

where T is the time corresponding to the chosen interval ($T = 0$, 1.5 or 3.0 sec for the early, middle or late interval, respectively). The parameter ρ scales the relative preference to earlier intervals, which captures an individual's temporal discounting or impatience to wait.

For each model, we fitted the model decision probabilities to the participant's choice interval data by minimizing the negative log-likelihood of the observed choices using Bayesian adaptive direct search (BADS) (Acerbi & Ma, 2017). Free parameters were optimised individually for each participant and separately for each block with the following boundaries: $\alpha \in [0,1]$, $\beta \in [0,20]$, $\rho \in [0,0.2]$, $w \in [0,1]$. The parameter w was fixed to 0 for the reinforcement learning model and fixed to 1 for the belief learning model. For the stochastic selection model, we fixed the parameters α and w to 0 so that this model could only capture the participants' decision uncertainty and their choice preference. This model produced constant model decision probabilities across all trials. To verify that we had found the global minimum, we repeated the search process with different starting points. For model comparison, we applied AICc—Akaike information criterion with a correction for finite sample size—to each participant and model as the information criterion for goodness-of-fit (Burnham & Anderson, 1998; Hurvich & Tsai, 1989). The summed AICc across participants was reported in Figure 5B.

Simulated play

During simulated play, an RL agent and a BL agent played against the prediction algorithm used by each class of competitor. The competitor's prediction was made using a history of choices a simulated agent made rather than using real data. For each simulated play of 60 trials (which is approximately equal to the number of trials in the real game), the success rate was computed. We simulated each agent's behaviour given a set of model parameters. Each set of parameters was determined by an extensive grid search in the parameters' space. The simulation play was repeated 3,000 times for each parameter set.

Latent change score model

Latent change score (LCS) models are the statistical framework that captures the process underlying the change in the variables of interest at two measurement occasions (Carpenter et al., 2019; Kievit et al., 2018; Kievit et al., 2017; McArdle, 2009). LCS models conceptualise the score of variable X at time point $T2$ as

$$X_{T2} = \beta X_{T1} + \Delta X$$

where the score X_{T2} is a function of the score X_{T1} weighted by an auto-regressive (i.e., self-feedback) parameter β and some residual ΔX. By fixing the regression weight of X_{T2} on X_{T1} to 1, the change score ΔX can be simply rewritten as

$$\Delta X = X_{T2} - X_{T1}$$

In structural equation modelling, the change score can be defined as a latent factor by fixing a factor loading on the score X_{T2} to 1. By this mathematical manipulation, the change between $T1$ and $T2$ is modelled as a latent factor. Bivariate LCS models predict the change score by an auto-regressive parameter β and a cross-coupling parameter α:

$$\Delta X = \beta X_{T1} + \alpha Y_{T1}$$

$$\Delta Y = \beta Y_{T1} + \alpha X_{T1}$$

In this equation, the auto-regressive parameter β captures the degree to which the initial score X_{T1} predicts (or is proportional to) the change score ΔX. The cross-coupling parameter α captures the degree to which the initial score in another domain Y_{T1} predicts (or is proportional to) the change score ΔX. Above these effects, the bivariate LCS models quantifies the variance-covariance structure in the change factor, which estimates the correlated change: the degree to which the change score in one domain ΔX covaries with the change score in another domain ΔY after taking auto-regressive and cross-coupling effects into account. See Kievit et al., 2018 & McArdle, 2009 for reviews and Kievit et al., 2017 & Carpenter et al., 2019 for its applications.

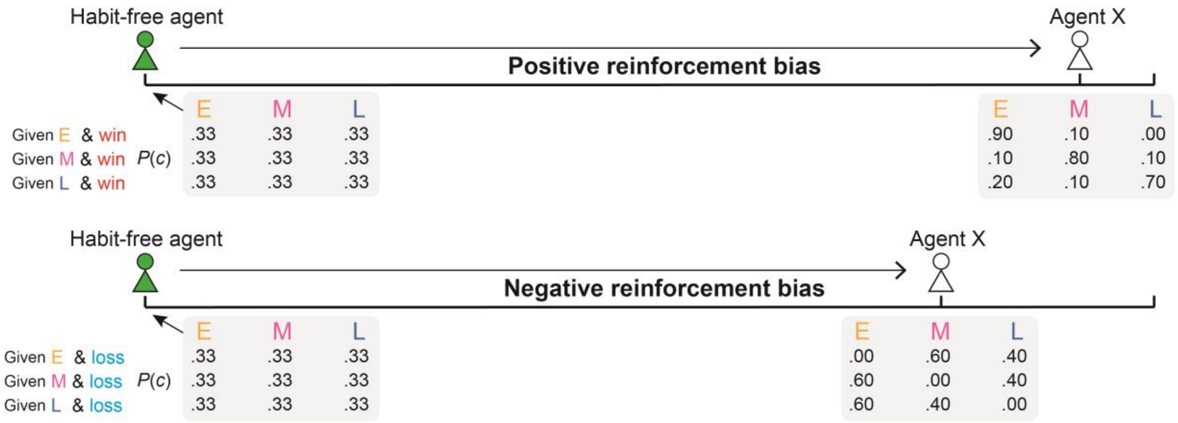
We examined the inter-relationships between adaptive autonomy and a shift in learning strategies. There were three decision bias scores (choice bias, sequential bias and reinforcement bias) and four hybrid model parameters (learning rate, decision uncertainty, decision preference and relative weight of belief learning). Considering potential correlations among seven variables, we controlled for the influences of the other five variables on bivariate changes. In Figure 6, we used the choice bias score and the belief learning weight. We first regressed all other five variables against these two scores and retaining only the residuals from the regression. We then used the residual scores in the choice bias and the residual scores in the belief learning weight at two measurement points to fit the bivariate LCS model. Therefore, any parameter estimates in the path model were considered a mere relationship between two variables included in the model. Because we fitted the bivariate LCS model to residuals, parameter estimates for the initial intercepts (i.e., mean initial scores) and change intercepts (i.e., mean change scores) were fixed to zero.

Models were estimated in the lavaan package for R (version 0.6-11) (Rosseel, 2012). We used maximum likelihood estimation with robust (Huber-White) standard errors and a scaled test statistic. We evaluated overall model fit using the root-mean-square error of approximation (RMSEA; acceptable fit: < 0.08; good fit < 0.05), the comparative fit index (CFI; acceptable fit: 0.95 to 0.97; good fit > 0.97) and the standardized root-mean-square residual (SRMR; acceptable fit: 0.05 to 0.10, good fit: < 0.05) (Schermelleh-Engel et al., 2003).

Supplementary information

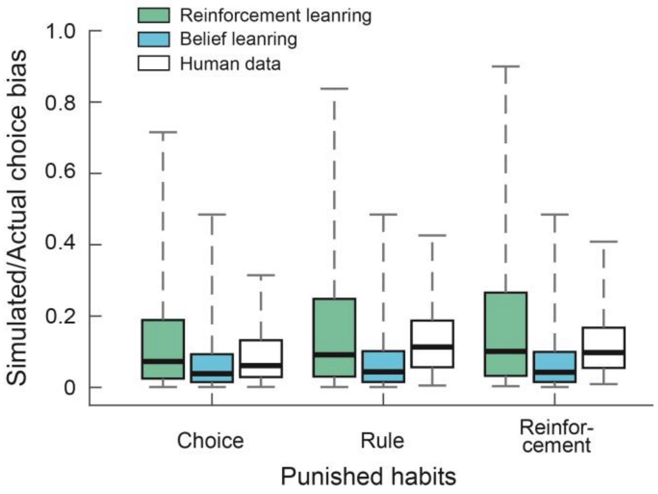
Exclusion criterion

We encouraged participants to sustain the percentage of timeout trials under 5%. We checked the histogram of the timeout rates. Seven participants displayed a timeout above 13 %. This was considerably high compared with the other participants (0–5%: 119 participants; 5–8%: 27 participants; 8–11%: 6 participants; 13–20%: 5 participants; >20%: 2 participants). These participants might not be able to follow the instructions or might not be able to keep their attention on the task, thereby removed from the analysis.


Habit-free agent
Agent X
Positive reinforcement bias
E M L
Given E & win .33 .33 .33
Given M & win P(c) .33 .33 .33
Given L & win .33 .33 .33
E M L
.90 .10 .00
.10 .80 .10
.20 .10 .70
Habit-free agent
Agent X
Negative reinforcement bias
E M L
Given E & loss .33 .33 .33
Given M & loss P(c) .33 .33 .33
Given L & loss .33 .33 .33
E M L
.00 .60 .40
.60 .00 .40
.60 .40 .00

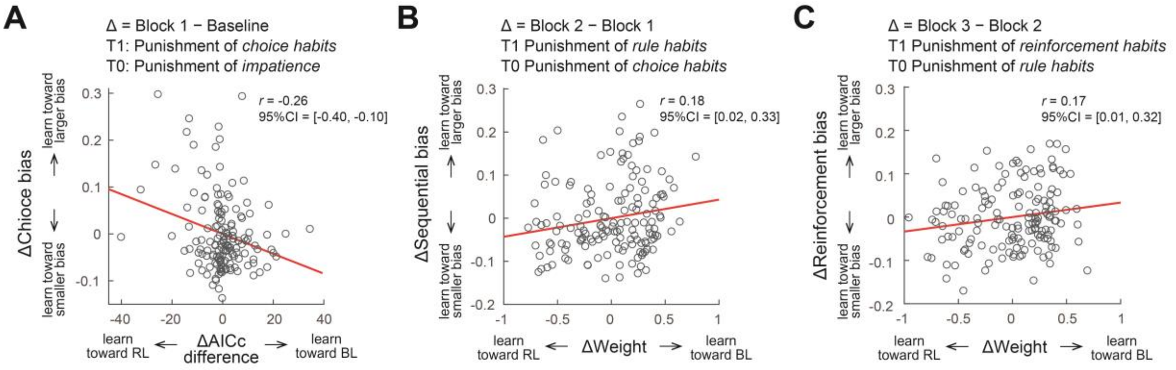
Supplementary Figure 1

The scores for positive reinforcement bias and negative reinforcement bias were computed from the conditional probabilities of wins only and from the conditional probabilities of losses only. These bias scores measure the statistical distance from random probabilities 0.33. See Figure 3.



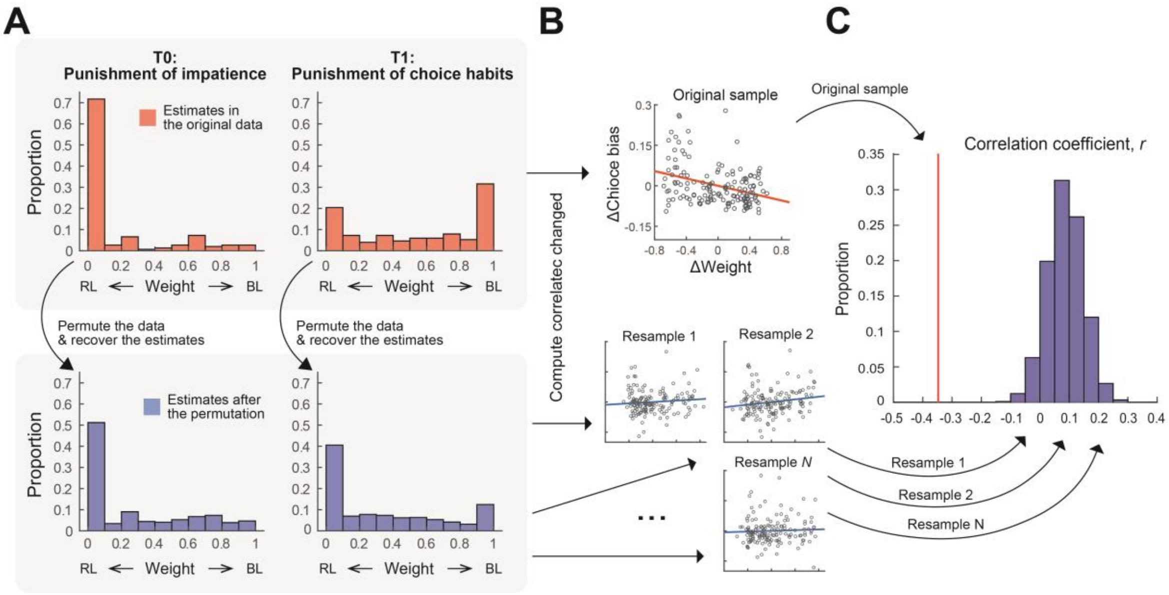
Supplementary Figure 2

Real choice bias (white bars) in the actual experiment and fictive choice bias (green or blue bars) in simulated play. We simulated the choice bias from agents using the reinforcement learning (green) and agents using the belief learning (blue) (see *Simulated play*). Belief learning produced a smaller choice bias in all punishment schemes.



Supplementary Figure 3

As supplementary results, we fitted the bivariate LCS model to the data set in panels **A**-**C**, respectively. Estimated parameters in the path model were omitted for the sake of space. Instead, we visualised the scatter plot of correlated change as we did in Figure 6B. **A**. Here we used the choice bias score and the AICc difference between the RL only model and BL only model, instead of the BL weight estimated under the hybrid learning rule. To estimate the latent change factors, we used these observed scores on the punishment of impatience in the baseline and the punishment of standard choice habits in block 1. We found a negative correlated change (standardised coefficient = -0.256, 95% CI = [-0.398 -0.101]): shifting the strategy toward belief learning as opposed to reinforcement learning was associated with gaining greater reductions in the choice bias. **B**. Here we used the scores of sequential bias and the scores of the BL weight on the punishment of choice habits in block 1 and the punishment of rule habits in block 2, to estimate the latent change factors. We did not find a negative correlated change (standardised coefficient = 0.180, 95% CI = [0.021 0.330]): shifting the strategy toward belief learning was not associated with gaining greater reductions in the sequential bias. **C**. Here we used the scores of reinforcement bias and the scores of the BL weight on the punishment of rule habits in block 2 and the punishment of reinforcement habits in block 3, to estimate the latent change factors. We did not find a negative correlated change (standardised coefficient = 0.168, 95% CI = [0.009 0.319]): shifting the strategy toward belief learning was not associated with gaining greater reductions in the reinforcement bias.



Supplementary Figure 4

A simulated experiment using permutation operation. **A**. In upper panels (red bars), we show the proportion of the belief learning weight estimated under the hybrid learning rule. A weight increases between blocks. In this model fitting, a trial sequence of rewards (success or failure) and that of birds' flight interval (early, middle or late interval) were used to simulate the participant's choice interval (see *Computational models*). If the competitor's action is critical for the model fitting, permutating the competitor's action would disrupt the estimates of the BL weight. To validate this, we swapped the birds' flight interval. For instance, if the participant chose the late interval and the birds intercepted the middle interval, we swapped the birds' choice to the early interval on that trial. This permutation does not alter an outcome (i.e., a successful trial is still success) but does randomise the competitor's action. We did not swap the birds' choice for unsuccessful trials. In each iteration, we permutated the birds' flight interval 75% of the time and estimated the BL weight under the hybrid rule. We repeated this procedure 100 times for each participant and for each block. In lower panels (blue bars), the parameter estimates recovered from the permutation sample are shown. The BL weights are right-skewed and indeed differ from the original estimates. **B**. In the upper panel, we show a scatter plot of the correlated change: how reductions in the choice bias co-occurs with a shift in learning strategies, replotted from Figure 6B. The correlation coefficient in this original sample (*r* = -0.347) is plotted as a vertical red line in the panel **C**. In lower panels, we plot the same correlated change but derived from the permutation sample. We only show three representative plots. In each sample, we computed the standardised coefficient of the correlated change. The proportion of these correlation coefficients is shown as a blue histogram in the panel **C**. The 95% confidence interval ranged from -0.03 to 0.20. The original correlation coefficient is significantly different from the permutated coefficients. This analysis validates the robustness of estimating the belief learning weight and the robustness of estimating the coefficient in the correlated change.

Competing interests

The authors declare no competing interests.

Author's contributions

Conceptualization, KO and PH; data collection, KO; investigation, KO, LC, and PH; formal analysis, KO; writing – original draft, KO; writing – review & editing, KO, LC and PH; supervision, LC and PH; funding acquisition, PH.

Acknowledgments

This research was supported by the John Templeton Foundation and the Fetzer Foundation awarded to PH.

Data and code availability

All data and analysis codes are available here:

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in neural information processing systems, 30*, 1834-1844.

Amodio, D. M., & Frith, C. D. (2006, Apr). Meeting of minds: the medial frontal cortex and social cognition. *Nat Rev Neurosci, 7*(4), 268-277. https://doi.org/10.1038/nrn1884

Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020, Feb). Gorilla in our midst: An online behavioral experiment builder. *Behav Res Methods, 52*(1), 388-407. https://doi.org/10.3758/s13428-019-01237-x

Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998, Nov). Random generation and the executive control of working memory. *Q J Exp Psychol A, 51*(4), 819-852. https://doi.org/10.1080/713755788

Baddeley, A. D. (1966, May). The capacity for generating information by randomization. *Q J Exp Psychol, 18*(2), 119-129. https://doi.org/10.1080/14640746608400019

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics, 12*(4), 428-454. https://doi.org/10.1016/0196-8858(91)90029-i

Barandiaran, X. E., & Di Paolo, E. A. (2014). A genealogical map of the concept of habit. *Front Hum Neurosci, 8*, 522. https://doi.org/10.3389/fnhum.2014.00522

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004, Apr). Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci, 7*(4), 404-410. https://doi.org/10.1038/nn1209

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15. https://doi.org/10.1016/0010-0277(94)90018-3

Brass, M., & Haggard, P. (2007, Aug 22). To do or not to do: the neural signature of self-control. *J Neurosci, 27*(34), 9141-9145. https://doi.org/10.1523/JNEUROSCI.0924-07.2007

Braver, T. S. (2012, Feb). The variable nature of cognitive control: a dual mechanisms framework. *Trends Cogn Sci, 16*(2), 106-113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). *Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control.* . Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780195168648.003

Burnham, K. P., & Anderson, D. R. (1998). Model Selection and Multimodel Inference. *Springer, New York, NY*.

Camerer, C., & Ho, T. (1999). Experience-weighted Attraction Learning in Normal Form Games. *Econometrica, 67*(4), 827-874. https://doi.org/10.1111/1468-0262.00054

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton Univ Press.

Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019, Jan). Domain-general enhancements of metacognitive ability through adaptive training. *J Exp Psychol Gen, 148*(1), 51-64. https://doi.org/10.1037/xge0000505

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007, May 29). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos Trans R Soc Lond B Biol Sci, 362*(1481), 933-942. https://doi.org/10.1098/rstb.2007.2098

de Leeuw, J. R. (2015, Mar). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav Res Methods, 47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

Dolan, R. J., & Dayan, P. (2013, Oct 16). Goals and habits in the brain. *Neuron, 80*(2), 312-325. https://doi.org/10.1016/j.neuron.2013.09.007

Du, Y., Krakauer, J. W., & Haith, A. M. (2022, May). The relationship between habits and motor skills in humans. *Trends Cogn Sci, 26*(5), 371-387. https://doi.org/10.1016/j.tics.2022.02.002

Dyson, B. J. (2021, Feb 3). Variability in competitive decision-making speed and quality against exploiting and exploitative opponents. *Sci Rep, 11*(1), 2859. https://doi.org/10.1038/s41598-021-82269-2

Dyson, B. J., Sundvall, J., Forder, L., & Douglas, S. (2018, Oct). Failure generates impulsivity only when outcomes cannot be controlled. *J Exp Psychol Hum Percept Perform, 44*(10), 1483-1487. https://doi.org/10.1037/xhp0000557

Fleming, S. M., Mars, R. B., Gladwin, T. E., & Haggard, P. (2009, Oct). When the brain changes its mind: flexibility of action selection in instructed and free choices. *Cereb Cortex, 19*(10), 2352-2360. https://doi.org/10.1093/cercor/bhn252

Forder, L., & Dyson, B. J. (2016, Sep 23). Behavioural and neural modulation of win-stay but not lose-shift strategies as a function of outcome value in Rock, Paper, Scissors. *Sci Rep, 6*, 33809. https://doi.org/10.1038/srep33809

Frith, C. (2013, Sep). The psychology of volition. *Exp Brain Res, 229*(3), 289-299. https://doi.org/10.1007/s00221-013-3407-6

Gehring, W. J., & Willoughby, A. R. (2002, Mar 22). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295*(5563), 2279-2282. https://doi.org/10.1126/science.1066893

Haggard, P. (2008, Dec). Human volition: towards a neuroscience of will. *Nat Rev Neurosci, 9*(12), 934-946. https://doi.org/10.1038/nrn2497

Haggard, P. (2019, Jan 4). The Neurocognitive Bases of Human Volition. *Annu Rev Psychol, 70*, 9-28. https://doi.org/10.1146/annurev-psych-010418-103348

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006, Feb). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biol Psychol, 71*(2), 148-154. https://doi.org/10.1016/j.biopsycho.2005.04.001

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008, May 6). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc Natl Acad Sci U S A, 105*(18), 6741-6746. https://doi.org/10.1073/pnas.0711099105

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*(2), 297-307. https://doi.org/10.1093/biomet/76.2.297

Jahanshahi, M., Dirnberger, G., Fuller, R., & Frith, C. D. (2000, Dec). The role of the dorsolateral prefrontal cortex in random number generation: a study with positron emission tomography. *Neuroimage, 12*(6), 713-725. https://doi.org/10.1006/nimg.2000.0647

Jahanshahi, M., Jenkins, I. H., Brown, R. G., Marsden, C. D., Passingham, R. E., & Brooks, D. J. (1995, Aug). Self-initiated versus externally triggered movements. I. An investigation using measurement of regional cerebral blood flow with PET and movement-related potentials in normal and Parkinson's disease subjects. *Brain, 118 (Pt 4)*, 913-933. https://doi.org/10.1093/brain/118.4.913

Jenkins, I. H., Jahanshahi, M., Jueptner, M., Passingham, R. E., & Brooks, D. J. (2000, Jun). Self-initiated versus externally triggered movements. II. The effect of movement predictability on regional cerebral blood flow. *Brain, 123 (Pt 6)*, 1216-1228. https://doi.org/10.1093/brain/123.6.1216

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A. L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Consortium, N., Lindenberger, U., & Dolan, R. J. (2018, Oct). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Dev Cogn Neurosci, 33*, 99-117. https://doi.org/10.1016/j.dcn.2017.11.007

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017, Jun). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood *Psychological Science 28*(10), 1419-1431. https://doi.org/https://doi.org/10.1177/0956797617710785

Kuroki, D. (2021, Feb). A new jsPsych plugin for psychophysics, providing accurate display duration and stimulus onset asynchrony. *Behav Res Methods, 53*(1), 301-310. https://doi.org/10.3758/s13428-020-01445-w

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavio*r (pp. 112–131). *Wiley, New York*.

Lee, D., Conroy, M. L., McGreevy, B. P., & Barraclough, D. J. (2004, Dec). Reinforcement learning and decision making in monkeys during a competitive game. *Brain Res Cogn Brain Res, 22*(1), 45-58. https://doi.org/10.1016/j.cogbrainres.2004.07.007

Lee, D., McGreevy, B. P., & Barraclough, D. J. (2005, Oct). Learning and decision making in monkeys during a rock-paper-scissors game. *Brain Res Cogn Brain Res, 25*(2), 416-430. https://doi.org/10.1016/j.cogbrainres.2005.07.003

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002, Jun). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *J Exp Psychol Appl, 8*(2), 75-84. https://doi.org/10.1037//1076-898x.8.2.75

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983, Sep). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain, 106 (Pt 3)*, 623-642. https://doi.org/10.1093/brain/106.3.623

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annu Rev Psychol, 60*, 577-605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron, 38*(2), 339-346. https://doi.org/DOI: 10.1016/s0896-6273(03)00154-5

Mischel, W., Ebbesen, E. B., & Zeiss, A. R. (1972, Feb). Cognitive and attentional mechanisms in delay of gratification. *J Pers Soc Psychol, 21*(2), 204-218. https://doi.org/10.1037/h0032198

Neiman, T., & Loewenstein, Y. (2011, Dec 6). Reinforcement learning in professional basketball players. *Nat Commun, 2*, 569. https://doi.org/10.1038/ncomms1580

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004, Apr 16). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452-454. https://doi.org/10.1126/science.1094285

Ota, K., Shinya, M., & Kudo, K. (2016, Nov 21). Sub-optimality in motor planning is retained throughout 9 days practice of 2250 trials. *Sci Rep, 6*, 37181. https://doi.org/10.1038/srep37181

Ota, K., Shinya, M., Maloney, L. T., & Kudo, K. (2019, Oct 16). Sub-optimality in motor planning is not improved by explicit observation of motor uncertainty. *Sci Rep, 9*(1), 14850. https://doi.org/10.1038/s41598-019-50901-x

Ota, K., Tanae, M., Ishii, K., & Takiyama, K. (2020, Jan 22). Optimizing motor decision-making through competition with opponents. *Sci Rep, 10*(1), 950. https://doi.org/10.1038/s41598-019-56659-6

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014, Jan 1). Bayesian model selection for group studies - revisited. *Neuroimage, 84*, 971-985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Robbins, T. W., & Costa, R. M. (2017, Nov 20). Habits. *Curr Biol, 27*(22), R1200-R1206. https://doi.org/10.1016/j.cub.2017.09.060

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007, Aug). The problem of serial order in behavior: Lashley's legacy. *Hum Mov Sci, 26*(4), 525-554. https://doi.org/10.1016/j.humov.2007.04.001

Rosseel, Y. (2012). lavaan: AnRPackage for Structural Equation Modeling. *Journal of Statistical Software, 48*(2). https://doi.org/10.18637/jss.v048.i02

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23–74.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009, Jul 15). Bayesian model selection for group studies. *Neuroimage, 46*(4), 1004-1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Story, G. W., Vlaev, I., Seymour, B., Darzi, A., & Dolan, R. J. (2014). Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective. *Front Behav Neurosci, 8*, 76. https://doi.org/10.3389/fnbeh.2014.00076

Sundvall, J., & Dyson, B. J. (2022). Breaking the bonds of reinforcement: Effects of trial outcome, rule consistency and rule complexity against exploitable and unexploitable opponents. *PLoS One, 17*(2), e0262249. https://doi.org/10.1371/journal.pone.0262249

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An Introduction, 2nd edition. *MIT Press, Cambridge*.

Tang, R., Bugg, J. M., Snijder, J. P., Conway, A. R., & Braver, T. S. (2022, Aug 10). The Dual Mechanisms of Cognitive Control (DMCC) project: Validation of an online behavioural task battery. *Q J Exp Psychol (Hove)*, 17470218221114769. https://doi.org/10.1177/17470218221114769

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., & Karpova, A. Y. (2014, Sep 25). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell, 159*(1), 21-32. https://doi.org/10.1016/j.cell.2014.08.037

Vickery, T. J., Chun, M. M., & Lee, D. (2011, Oct 6). Ubiquity and specificity of reinforcement signals throughout the human brain. *Neuron, 72*(1), 166-177. https://doi.org/10.1016/j.neuron.2011.08.011

Wang, Z., Xu, B., & Zhou, H. J. (2014, Jul 25). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Sci Rep, 4*, 5830. https://doi.org/10.1038/srep05830

Wong, A., Merholz, G., & Maoz, U. (2021, Oct 19). Characterizing human random-sequence generation in competitive and non-competitive environments using Lempel-Ziv complexity. *Sci Rep, 11*(1), 20662. https://doi.org/10.1038/s41598-021-99967-6

Wood, W., & Runger, D. (2016). Psychology of Habit. *Annu Rev Psychol, 67*, 289-314. https://doi.org/10.1146/annurev-psych-122414-033417

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013, Apr). Heterogeneity of strategy use in the Iowa gambling task: a comparison of win-stay/lose-shift and reinforcement learning models. *Psychon Bull Rev, 20*(2), 364-371. https://doi.org/10.3758/s13423-012-0324-9

Zhu, L., Mathewson, K. E., & Hsu, M. (2012, Jan 31). Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning. *Proc Natl Acad Sci U S A, 109*(5), 1419-1424. https://doi.org/10.1073/pnas.1116783109