Autonomous behaviour and the limits of human volition

Keiji Ota[1], Lucie Charles[1,2], Patrick Haggard[1]

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

[2] Department of Psychology, Queen Mary University of London, London, United Kingdom

Corresponding author

Keiji Ota Email: keiji.ota@ucl.ac.uk

Institute of Cognitive Neuroscience

University College London

17 Queen Square, London, WC1N 3AZ

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, autonomy and freedom imply unconstrained behaviour initiated by endogenous triggers. 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 external triggers and of routine response patterns, but nevertheless show appropriate coordination with the environment. Participants (N = 152) 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 (such as always responding early), sequential patterns of action timing across trials (such as early, late, early, late…), or predictable action-outcome dependence (such as win-stay, lose-shift). Adaptive autonomy was quantified as the ability to maintain performance when each of these influences on action selection was punished. We found that participants could become free from habitual choices regarding when to act and could also become free from sequential action patterns. However, they were not able to free themselves from influences of action-outcome dependence, even when these resulted in poor performance. These results point to a new concept of autonomous behaviour as flexible adaptation of voluntary action choice under different classes of internal constraint.

Keywords

Cognitive control / Free action / Habitual behaviour / Competitive pressure / Voluntary action

Introduction

The capacity for voluntary action is often considered a defining feature of the human mind, but what volition is remains controversial. Two strong aspects of the definition of volition rely on exclusion: actions that are immediate responses to an external triggering stimulus are less voluntary than actions that are internally generated; actions that merely continue a routine, stereotyped generative pattern are less voluntary than actions that are planned a new, or actions that arise spontaneously. Both of these key features of volition are lacking in laboratory studies of behaviour, which traditionally seek tight experimental control through imperative stimuli and overlearned response patterns.

Many studies of volition involve a paradoxical instruction to act in a way that is endogenous and spontaneous (Brass & Haggard, 2007; Fleming et al., 2009; Jahanshahi et al., 1995; Libet et al., 1983) without being triggered by external factors (Jenkins et al., 2000). This minimalist approach can elicit actions that are stimulus-independent. Often, it leads to actions that have an unpredictable or random pattern, either because this is explicitly instructed (Baddeley et al., 1998; Baddeley, 1966; Jahanshahi et al., 2000), or because the participant implicitly intuits that stereotypy and voluntariness are opposites. But such studies neglect the reasons or goals that an agent may have for acting, and omit the crucial 'Why?' aspect of voluntary action (Haggard, 2019; Shadlen & Roskies, 2012).

Philosophical accounts of free action distinguish between freedom **from** external constraint, and freedom **to** act in accordance with goals and desires. Here we propose that competitive game scenarios can capture both the "freedom from" aspect of voluntary action (since the game structure can require the participant to act in the absence of triggering stimuli, or stereotyped response patterns) and "freedom to" aspect (since the participant has the clear goal of defeating the opponent and achieving good performance). For example, in the 'rock, paper, scissors' game, each participant selects an action without first seeing the action of their opponent. Second, many competitive games encourage unpredictable action over stereotyped action, even without any explicit instruction to act randomly. In addition, many games have a "means-ends" structure, where active selection between alternative possible action choices is the key to success. In order to defeat the opponent, each participant needs to leave from their stereotyped exploitative behaviours and instead explore, innovate a new strategy that possibly brings them to successful performance (thus monetary rewards). Humans and non-human primates indeed respond to competitive pressure by initiating exploratory behaviour (Barraclough et al., 2004; Lee et al., 2004; Lee et al., 2005).

In the present study, we have studied a classical paradigm of volition through the lens of a competitive game. Participants chose when to make a simple manual keypress action (Libet et al., 1983), while trying to avoid colliding with a competitor. We designed several virtual competitor algorithms, each designed to predict and punish stereotyped choice patterns associated with a particular kind of cognitive strategy for action generation, pressurising players to become free from the corresponding pattern. The ability to adapt voluntary action choices under various competitive pressures offers quantitative measures of individual autonomy. These measures can identify how free the agent can become from specific action-generation strategies or cognitive habits and how successful they are in remaining free to achieve their goal (of succeeding in the game). For example, if people are unable to avoid repeating the same action choice, even when the current game environment punishes them for doing so and thus providing good reasons to innovate actions, one might question in what sense their actions are truly free. This conceptualization of free and autonomous action is widely used in understanding addiction (Weinberg, 2020; Weinberg, 2022).

We conceptualised three distinct expressions of autonomous behaviour, corresponding to three different constraints from which their action should become free. First, people may have a bias towards simple repetition of a given action choice (Dolan & Dayan, 2013; Robbins & Costa, 2017). We refer to this pattern of behaviour as *choice bias*. Consider the simple task of repeatedly generating one of three digits (see Figure 1A). Agent X may prefer to choose “1”, for whatever reason. Suppose now that a competitor punishes X for repeating one choice within a game scenario. If X’s action is reasons-responsive to the environmental challenge, agent X should now choose the two other digits more often and should approach an agent who has no preference on any options. This adaptive capacity may reflect the autonomy each agent has over their choice bias. We call this quantity *adaptive autonomy*.

In the second environment, the competitor pressurised participants to change any habitual action chains or routines (Lashley, 1951; Robbins & Costa, 2017; Rosenbaum et al., 2007). In our example, integer counting (“1, 2, 3”) is such a choice pattern (agent X in Figure 1B). We refer to this as *sequential bias*. The only way to completely avoid such biases 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; Ginsburg & Karpiuk, 1994). Figure 1B illustrates potential behavioural adaptation to punishing sequential patterns.

Lastly, we consider how people respond to previous action successes and failures, by considering *reinforcement bias*. In Figure 1C, agent X repeats the digit that has been rewarded.

In contrast, we can imagine an agent who generates each choice independently from whether the previous outcome was rewarded or not. Most reinforcement learning approaches 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 pattern of outcome-dependent action choices when it is punished by a competitor. Adaptive autonomy would mean that an agent would be able to break the association between the previous outcome and the forthcoming action (Fig. 1C). Such freedom from influences of past reinforcement is considered important in situations such as combatting addiction, and recovery from trauma.

In addition, we considered whether adaptive autonomy with respect to these three different forms of constraint could be controlled by a shared cognitive mechanism (Braver, 2012; Braver et al., 2007; Tang et al., 2022), or whether there are distinct cognitive control modules specific to each particular bias. To do this, we explored correlations across individuals in our adaptive autonomy measures. Finally, we modelled the learning process by which people generated a new action in order to avoid colliding with the competitor. Due to the nature of the competitive game, the participant could potentially develop a model of the opponent by understanding what options the opponent is going to select, and thus improve their beliefs about the opponent’s behaviour and guide their own action choices accordingly. We thus examined whether participants moved from model-free reinforcement learning (Dolan & Dayan, 2013) towards model-based behaviour. This would represent an interesting convergence between theory of mind, voluntary action, and behavioural autonomy.

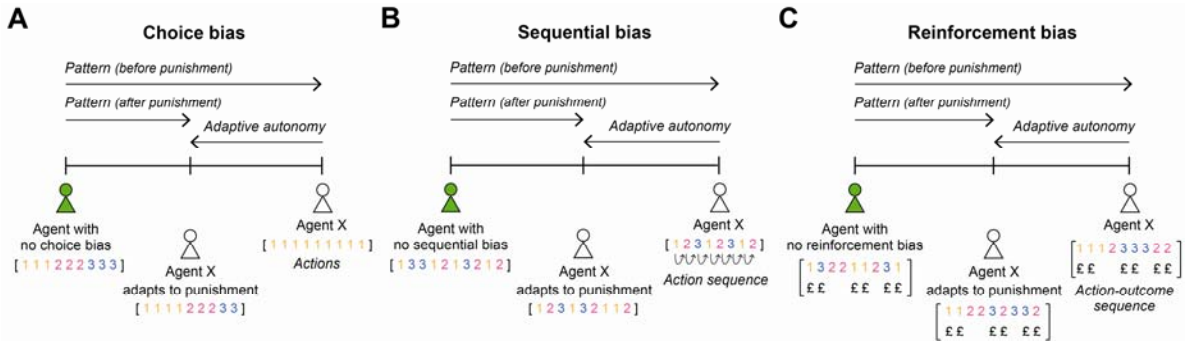

Figure 1. Hypothetical experiment. An agent is required to generate digits 1, 2, 3, as shown in the square brackets. **A**. Agent X exclusively selects the digit 1. They have a choice bias. In contrast, an agent with no choice bias selects each digit equally often. **B.** Agent X generates successive digits by counting. In contrast, an agent with no sequential bias does not use any

obvious transition strategy between moves. **C.** Agent X tends to repeat any action that has just been rewarded. In contrast, an agent with no reinforcement bias does not use information about the outcome of the present action to decide their next action. In each case, the leftward arrow indicates the reduction in a given bias when the opponent punishes that pattern of action choice by using it to predict the agent's next move, i.e., adaptive autonomy.

Methods

The full description of the data analysis, computational models and structural equation modelling can be seen in Supplementary Material.

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 due to insufficient performance 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). A short version of our task is available to play online: https://research.sc/participant/login/dynamic/5D39406C-8AB9-4F6D-8FD7-9F995D9DCB83

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 and pass through the storage location. 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. 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 moment when a stimulus is absent – 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 tends to act. Therefore, predicting when the birds will likely fly out of the tree and randomising the time to throw are important. The participants were not told explicitly about the strategy that the competitor used.

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 remaining time 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%.

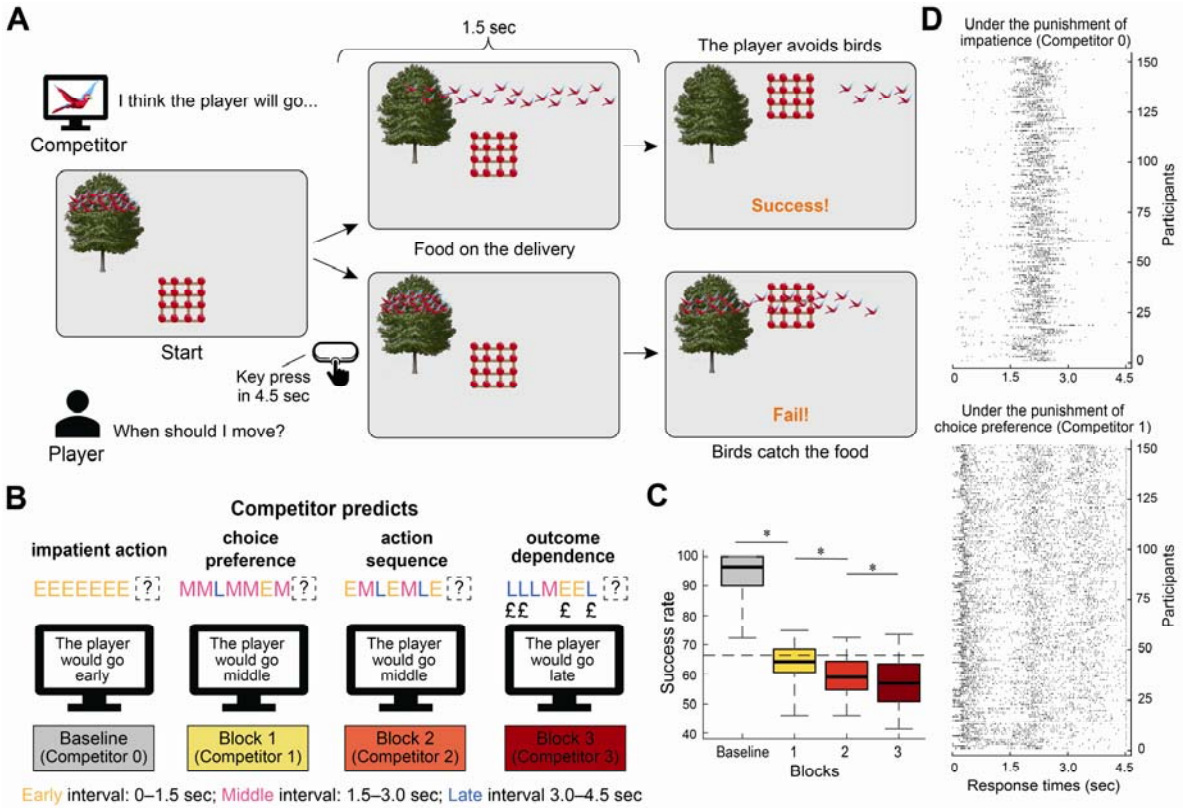

Figure 2. Virtual competitive environments. **A**. A trial sequence. A participant decides when to throw food within a 4.5 second time window. The food is delivered at the top-centre of the screen 1.5 sec after their key press. A virtual competitor (i.e., a flock of birds) attempts to intercept the food by adjusting the time at which it flies out of a tree. Participants win a trial if they throw at a time which avoids the food being caught by the birds. **B**. Experimental design. On each trial, the virtual competitor intercepted the food during one of three action intervals, 1)

early throw (0–1.5 sec), middle throw (1.5–3.0 sec) or late throw (3.0–4.5 sec). In the baseline block, the early interval, associated with impatience, was punished. In block 1, any *choice preferences* that favoured any one interval over all others were predicted and punished. In block 2, any *sequential action patterns* were punished. In block 3, any *outcome dependences* were punished. **C**. Success rate of avoiding the birds against each type 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 during the punishment of impatience in the baseline block (upper panel) and during the punishment of choice preference in block 1 (lower panel). The response time data for each participant are aligned in each row. Each small dot represents a reaction time in each trial.

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 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 *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 *sequential action 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. Competitor 2 then exploited these conditional probabilities to predict the upcoming choice given the last reaction time. Competitor 2 pressured participants in avoiding habitual transition from one action to the next.

Finally, in block 3, Competitor 3 punished *outcome dependences* – 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 looked for the conditional probabilities of selecting three intervals to predict the upcoming choice given 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 decision bias scores. We quantified three different decision bias scores (choice bias, sequential bias, reinforcement bias) that each individual has in their action generation patterns in order to examine whether individuals reduced a particular bias score (e.g., choice bias) when a corresponding pattern (e.g., choice preference) was punished by the competitor. The extent to which individuals were able to reduce their decision bias score would indicate their *adaptive autonomy* with respect to a corresponding internal constraint.

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 bias-free agent would exhibit. See Figure 3.

1) Choice bias. The probabilities of choosing the early, middle and late interval for a bias-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

$$D_{KL\ choice\ bias} = \sum_{c\ \in\ E,M,L} P(c)\log_2\left(\frac{P(c)}{0.33}\right) \tag{1}$$

2) Sequential bias. 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

$$D_{KL\ c_{-1}} = \sum_{c\ \in\ E,M,L} P(c|c_{-1})\log_2\left(\frac{P(c|c_{-1})}{P(c)}\right) \tag{2}$$

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

$$D_{KL\ sequential\ bias} = \sum_{c_{-1}\ \in\ E,M,L} P(c_{-1})\cdot D_{KL\ c_{-1}} \tag{3}$$

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, on the top of choice preference where Competitor 1 punished.

3) Reinforcement bias. 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

$$D_{KL\ c_{-1},o_{-1}} = \sum_{c\ \in\ E,M,L} P(c|c_{-1}, o_{-1})\log_2\left(\frac{P(c|c_{-1}, o_{-1})}{P(c|c_{-1})}\right) \tag{4}$$

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

$$D_{KL\,reinforcement} = \sum_{c_{-1} \in E,M,L} \sum_{o_{-1} \in \substack{success \\ fail}} P(c_{-1}, o_{-1}) \cdot D_{KL\ c_{-1},o_{-1}} \tag{5}$$

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, on the top of choice preference and sequential pattern punished by Competitor 2.

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).

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 intercept and 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). Participants were not told whether the birds chose a continuous or a 3-option distribution. 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. Participants could not 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 were present and then threw in reaction to that moment, the birds could suddenly appear and intercept the food. Therefore, the participants were instructed to predict when the birds would appear and to try avoid them. This feature means that our participants' actions were stimulus-independent.

There were 4 blocks in total. In each block, the participants competed with a class of competitor that pressurised a distinct internal constraint or a stereotypical action-generation strategy (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 choice preference: if a participant selected one interval more often than the other two, then the birds became more likely to intercept the food during the same interval. In block 2, Competitor 2 predicted sequential pattern, and punished any predictable association between the time of the participant's current throw and the time of the preceding throw. Finally, in block 3, Competitor 3 punished outcome dependence 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 required progressive degrees of autonomy across blocks: they needed in each block to act in a way that was even more unconstrained than required by the competitors they had played previously.

Participants were never told when they should act on any given trial. Participants did not receive any explicit instruction or explanation about how the competitor would behave or respond to the participant in each particular block. 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 behavioural adaptation rather than explicit understanding.

We first checked 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 a mean 2.15 sec (SD 0.49 sec) after the trial started (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 increased the predictive power of punishing the participants' responses.

Do people adapt to punishments of habitual actions?

We next examined the extent to which people could adapt to punishment of different behavioural patterns. We therefore developed a measure that reflects an individual's tendency towards the pattern of action selection for punishment on each block. 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 an agent, who does not have a particular type of bias would exhibit (Figure 3A-C). A lower statistical distance means that the individual is similar to this bias-free agent. We call this quantity a *decision bias score*. The decision bias score *before* the punishment of a specific pattern reflects an individual's internal constraint regarding when to act. We then looked at the *change* in decision bias scores relative to the previous block, when a given individual pattern was punished. We quantified the score for choice bias, sequential bias and reinforcement bias, respectively (Figure 3A-C). A greater change in bias score would indicate stronger *adaptive autonomy*, or ability to update the reasons-guided voluntary action choice.

Participants began to distribute action times appropriately when Competitor 1 started punishing the repetition of the same action interval (Fig. 2D). Accordingly, their choice bias score—a statistical distance of the observed choice probabilities from probabilities 0.33 (Fig. 3A)—reduced after punishment (Fig. 3D; Mdn = 1.37 for the punishment of impatience versus Mdn = 0.06 for the punishment of choice preference, $p < .001$, $z = 10.69$, Wilcoxon sign rank). Competitor 1 did not seek sequential patterns from one action to the next and allowed participants to still use a rule-based action chain. We quantified the sequential bias score by considering the extent to which probabilities given the previous action are explained by one's own choice probabilities (Fig. 3B). We found that the sequential bias decreased after the competitor pressurised action sequence (Fig. 3E; Mdn = 0.13 for the punishment of choice preference versus Mdn = 0.09 for the punishment of action sequence, $p < .001$, $z = 3.82$, Wilcoxon sign rank). Against Competitor 3, the participants were asked to act even more freely to avoid patterns following positive or negative outcome. We evaluated the reinforcement bias score by considering the extent to which probabilities given the combination of previous action and the previous outcome are explained by probabilities given the previous action alone (Fig. 3C). The reinforcement bias did not show any significant improvement under competitor punishment (Fig. 3F; Mdn = 0.11 for the punishment of action sequence versus Mdn = 0.11 for the punishment of outcome dependence, $p = .79$, $z = 0.27$, Wilcoxon sign rank).

We further tested the possibility that the participants adapted differently to the influence of positive and negative reinforcement. The processes that follow a positive outcome may be

different from those that follow a negative outcome (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 score separately (SFig. 1A). Nevertheless, we did not find significant improvements in the positive reinforcement bias (SFig. 1B; Mdn = 0.29 for the punishment of action sequence versus Mdn = 0.34 for the punishment of outcome dependence, p = .21, z = -1.26, Wilcoxon sign rank) nor in the negative reinforcement bias (SFig. 1B; Mdn = 0.43 for the punishment of action sequence versus Mdn = 0.41 for the punishment of outcome dependence, p = .19, z = 1.30, Wilcoxon sign rank). These results suggest that people are able to become more autonomous from repeating the same action interval and habitual action transitions but cannot break away from outcome dependencies. That is, people display stereotyped patterns of being guided by reinforcement, such as win-stay lose-shift, even when they are discouraged to do so.

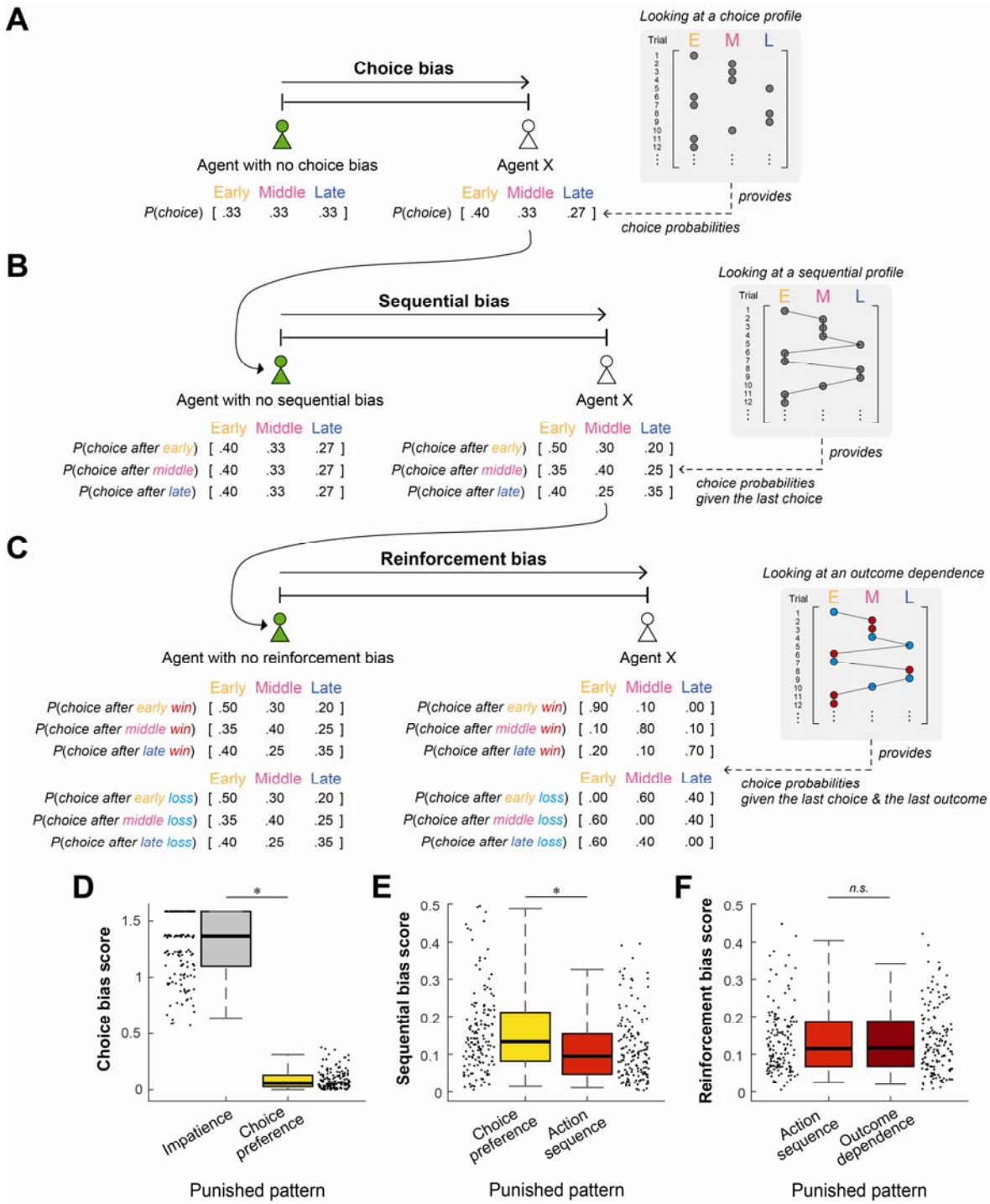

Figure 3. Measuring decision bias scores with respect to different punishments. In the raster plot, potential actions for the early, middle and late intervals are shown. A vector of the numerical values illustrates the probabilities for choosing these three actions. **A.** Agent X selects the early interval more frequently than the other two intervals. In contrast, an agent

without choice bias selects each interval equally often. We measured the statistical distance between individual and a bias-free agent with respect to a pattern similarity between their choice probabilities. **B.** The dashed line from the choice bias vector in A to the sequential bias vector in B shows the logical relation between choice and transition: any change in choice bias inevitably changes the transition vector. Sequential bias is considered as the additional change over and above such basic effects of the choice bias. An agent without any sequential bias should determine their next action independently from the previous action. **C.** The dashed line from the sequential bias vector in B to the reinforcement vector in C shows the logical relation between transitions and reinforcements. Reinforcement bias is considered as the additional change over and above such basic effects of sequential bias. An agent without any reinforcement bias should determine their next action independently from the previous success or failure. **D-E.** Planned comparisons showing the degree of adaptive autonomy (bias reduction) when choice preference is punished (D), when action sequence is punished (E) and when outcome dependence is punished (F). * 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. Dots adjacent to each box shows individual scores.

Is there a common factor underlying all adaptive autonomy?

The bias reduction between the pre-punishment and the post-punishment phase provides a measure of adaptive autonomy for each decision bias (Fig. 3D-F). Theories of domain-general cognitive control propose that proactive, strategic cognitive control acts similarly across cognitive tasks (Braver, 2012; Braver et al., 2007; Tang et al., 2022). The adaptive ability to become free of one kind of bias would be expected to correlate with the ability to become free of another bias, to the extent that both depend on a common, domain-general mechanism. In fact, the correlation structure of the changes in our decision bias scores showed poor correlations across our 152 participants (Figure 4; r = -0.07~-0.01). This suggests that the ability to voluntarily regulate one bias is not strongly associated with the ability to regulate other biases. 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. Since this correlation approach would be undermined by a floor effect of participants who showed no adaptive autonomy at all, we repeated the correlations including

only participants who showed a numerical change in decision bias. We did not find significant correlations among adaptive autonomy to the three bias classes even in this restricted group (r = -0.06~0.04). These results suggest that people recruit distinct cognitive capacities to exert autonomous behaviour when unlearning choice bias, sequential bias and reinforcement bias, contrary to concepts of a general cognitive control capacity.

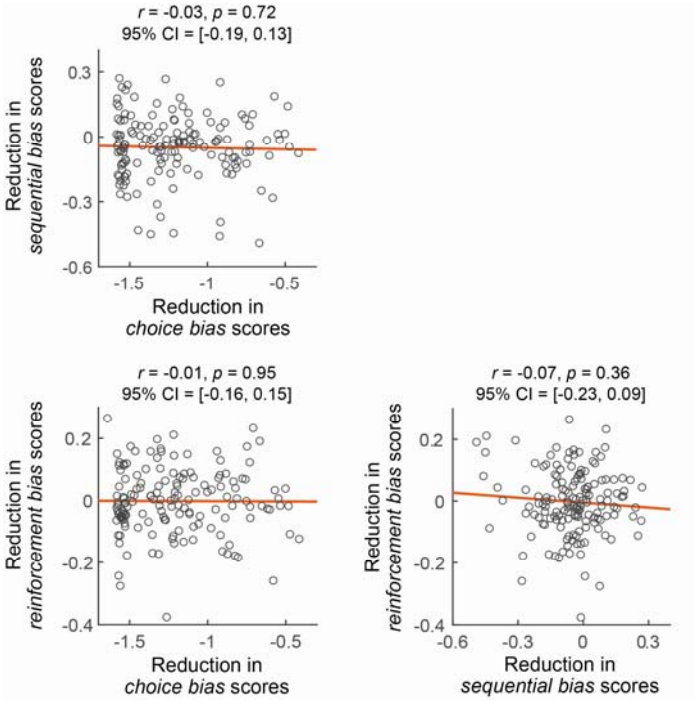

Figure 4. Correlation structure underlying measures of adaptive autonomy. A planned comparison between the pre-punishment and the post-punishment phase in Figure 3D-F provides bias reduction (i.e., adaptive autonomy) for each class. Across 152 participants, there is no strong correlation between the three measures of adaptive autonomy.

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 in which actions are chosen randomly. This makes behaviour unpredictable but involves no interaction with the 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 in this task is considered belief learning (BL). Here we consider whether our participants also employed BL by predicting when the birds would fly and avoiding them. If so, this would indicate that voluntary action choice was made not because a particular choice was rewarded in the past but because participants predicted that the birds are less likely to fly at a particular moment on the *current* trial.

BL goes beyond reinforcement learning (RL) (Sutton & Barto, 2018). If participants used only RL, they would update the value of the action interval they just made, based on whether the throw was successful or fail. In contrast, if participants used BL, they would form a view about which interval the birds were likely to fly in the next trial, and could therefore update the value of all action intervals. This difference in the internal processes allows a BL agent to update the values of the action intervals, thereby prompting a frequent update of the best action. We simulated agents using either RL or BL in our competitive game (see *Simulated play* in Supplementary Material), and found that belief learning indeed induces a smaller choice bias than simple reinforcement learning (SFig. 2), leading to better performance (Fig. 5A). 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 preference (Fig. 5A; Mdn = 60% for RL versus Mdn = 66.6% for BL) and under the punishment of action sequence (Fig. 5A; Mdn = 51.6% for RL versus Mdn = 63.3% for BL). BL is therefore potentially useful for acquiring adaptive autonomy under competitor pressure.

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* in Supplementary Material). Across 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 (baseline). However, the BL provided a better fit to our participants' action when the competitor started punishing choice preference. The BL model still outperformed when the competitor punished sequential pattern and outcome dependence (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 preference and action sequence, and this was 97% under outcome dependence (Fig. 5C).

We also fitted a hybrid learning rule which combines reinforcement learning and belief learning (Camerer & Ho, 1999; Hampton et al., 2008; Zhu et al., 2012). The hybrid learning rule fitted the data relatively well in all blocks (Fig. 5B). The hybrid learning rule has an additional parameter that describes the weight given to belief learning relative to reinforcement learning. The weight placed on BL in the hybrid model was strongly correlated with the model fit index under the BL only model relative to the RL only model (Fig. 5D). Across blocks, the weight parameter on BL increased from punishment of impatience to punishment of choice preference (Fig. 5E; Mdn = 0.00 for the punishment of impatience versus Mdn = 0.63 for the punishment of choice preference, $p < .001$, $z = 8.32$, Wilcoxon sign rank). That is, belief learning made an important contribution to participants' choices when they were punished for choice preference (Fig. 5E; Mdn = 0.63), action sequence (Mdn = 0.81; versus choice preference, $p = .10$, $z = 1.65$) and outcome dependence (Mdn = 0.74; versus action sequence, $p = .96$, $z = 0.04$). 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 action choice patterns, participants switched to learning successful actions from the opponent's prior actions. This learning process is considered model-based cognitive control in the sense that the likelihood of the opponent's action is updated and developed. A model comparison thus indicates that adaptive autonomy is related to a shift from model-free control towards model-based control, rather than simply behaving randomly. In the next structural equation modelling, we examine this statistical relationship between adaptive autonomy and belief learning.

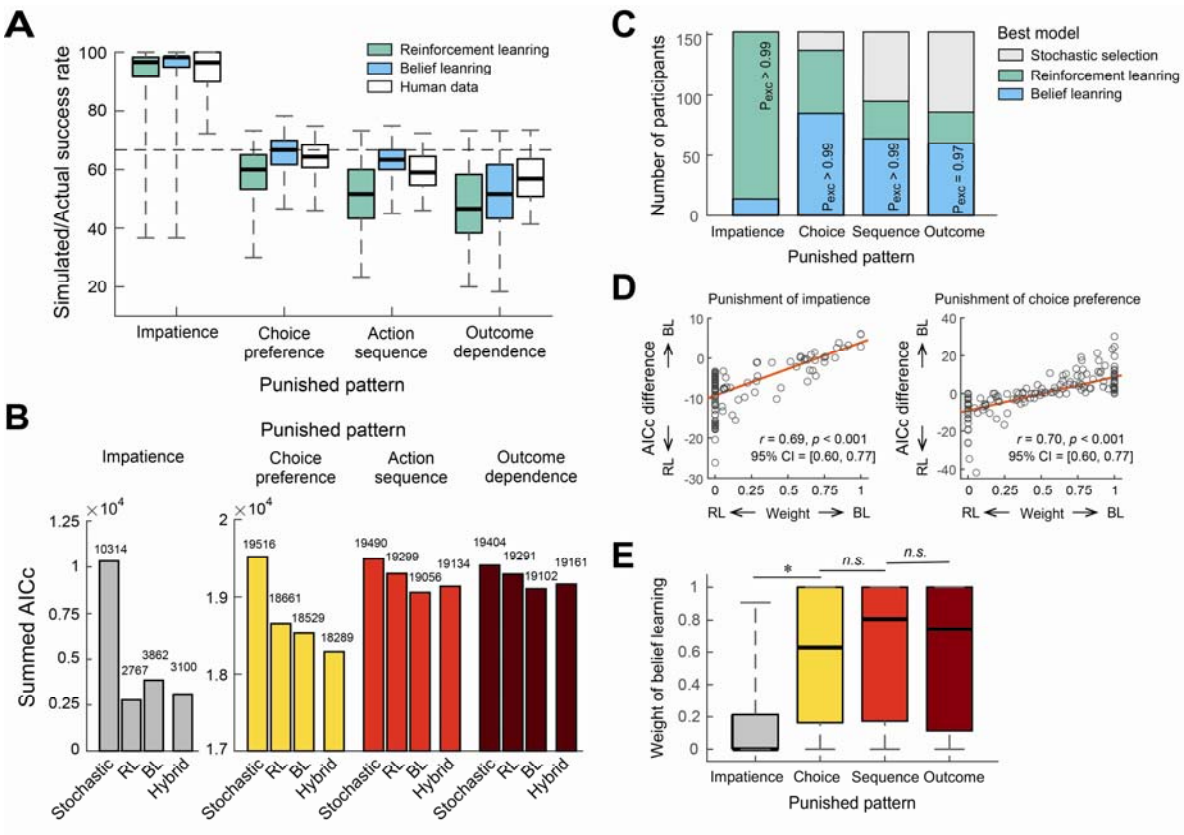

Figure 5. A shift in the strategy to belief learning as competitive demand increases. **A**. Real success rate (white bars) in the actual experiment and fictive success rate (green or blue bars). In simulation, agents who use the reinforcement learning strategy and agents who use the belief learning 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 fitted by the model. The P_{exc} inserted in the bar denotes the protected exceedance probability that supports the corresponding model. **D.** A weight placed on belief learning relative to reinforcement learning under the hybrid learning rule. As the weight increases, the BL only model provides a better fit than the RL only model. **E.** A transition in the belief learning weight across blocks. * 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

To this end, we used a 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* in Supplementary Material). By using a bivariate LCS model, two latent change factors can be modelled. This model allows us to estimate correlated changes between latent factors: the degree to which the bias reduction co-occurs with the change in weight placed on belief learning (i.e., a shift in learning strategies). If participants behaved simply stochastically and reduced a decision bias score, the bias reduction would not co-occur with the change in weight. If, on the other hand, participants adopted belief learning and avoided choice bias, the bias reduction would co-occur with an increase in BL weight.

To address this question, we fitted the bivariate LCS model to our data set. Our observed variables were the choice bias score and BL weight under the hybrid learning rule. To estimate the latent change factors, we used these two scores on the punishment of impatience in the baseline and on the punishment of choice preference in block 1, because our planned comparison focused on adaptive autonomy of choice bias 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 choice preference). 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). In bivariate LCS models, three kinds of paths can be estimated. The first is the extent to which the bias reduction is explained by the initial bias before punishment, which is termed auto-regression in LCS models. We found negative auto-regressive effects. This means that individuals who started off with a smaller bias in the baseline gained a smaller bias reduction in block 1, possibly because of a scale attenuation effect by reaching to the lower limit of the bias score. The second is the extent to which the bias reduction is explained by the initial BL weight, which is termed cross-coupling. There were no significant cross-coupling effects: a greater bias reduction in block 1 was not associated with individuals who started off with a greater weight in the baseline.

Notably, we found a medium negative correlated change. This suggests that individuals who gained a greater weight (i.e., a grater shift from RL to BL) between blocks also gained a greater bias reduction when choice preference was punished (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 validate this finding, we estimated the BL weight under simulation where the competitor's action was randomised. We found no correlated change between bias reduction and the change in the BL weigh (SFig. 3), suggesting that our participants likely used the information about the competitor's action. To summarise, these analyses demonstrate that adaptive autonomy is enhanced by developing the intention of the opponent's action through active interaction with a circumstance that is designed to evoke autonomous behaviour: a greater shift toward model-based control is related to boosted autonomy.

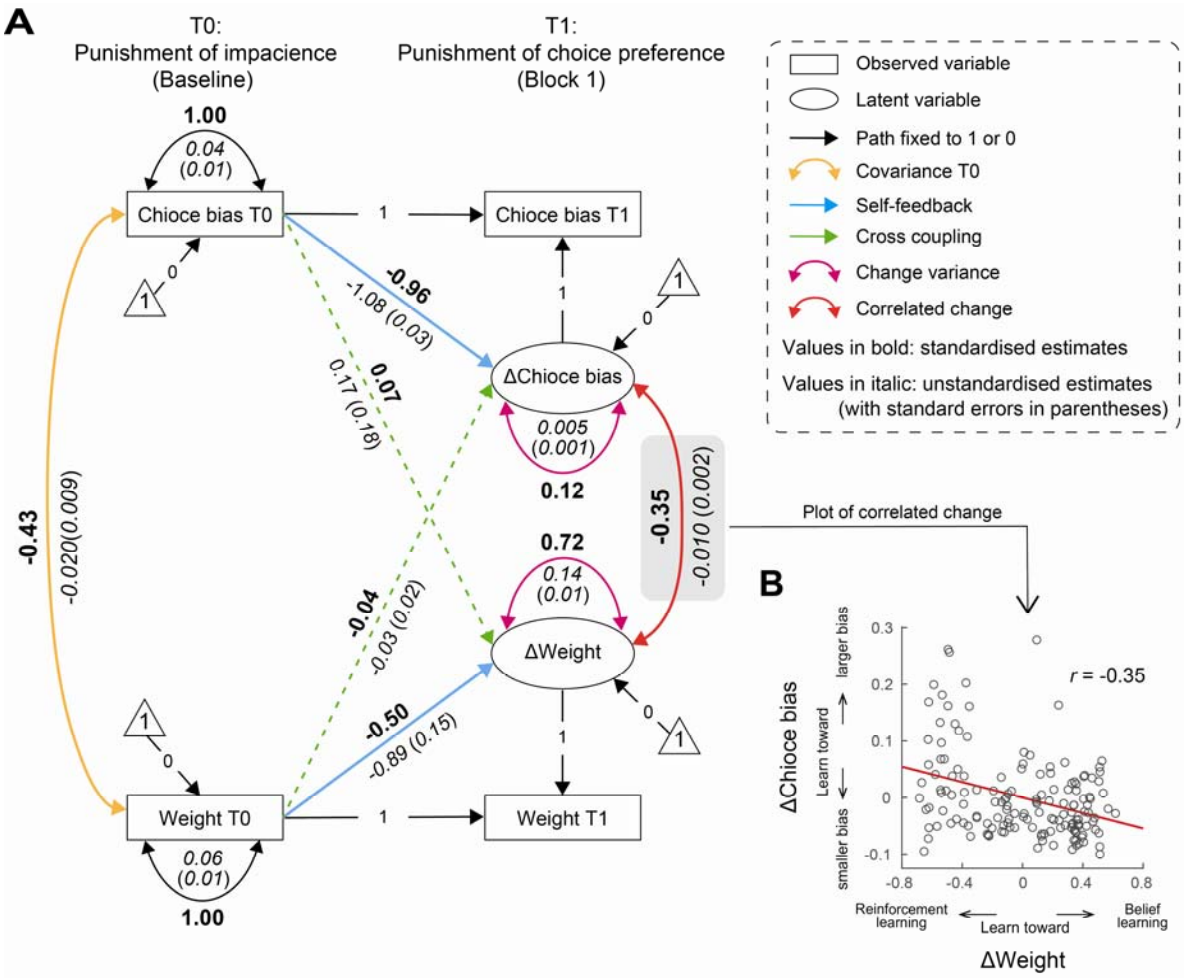

Figure 6. The bivariate latent change score model of adaptive autonomy and strategic learning. **A**. Estimated parameters. The change score in the choice bias was modelled as a latent factor between the score before the punishment of choice preference (baseline) and that under the punishment (block 1). A latent change in the belief learning weight 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. The reduction in choice bias score is correlated with a greater shift from reinforcement learning to belief learning.

Discussion

It is widely assumed that humans choose and control their own actions for reasons that are important to them. However, most experimental psychology struggles to highlight key features of human voluntariness in the laboratory since they paradoxically instruct participants to act voluntarily. Previous neuroscientific literature has emphasized that a volitional action must be endogenous and spontaneous, rather than stereotyped. Previous philosophical literature has emphasized that a volitional action is made for a reason. Human experimental studies have generally struggled to combine these two seemingly contradictory features of volition (Maoz & Sinnott-Armstrong, 2022; Roskies, 2010). Here we have developed a family of novel competitive games which allow participants to act in a way which is simultaneously endogenous, non-stereotyped, and reasons-responsive. In the game, participants had to endogenously generate a new pattern of behaviour (without any immediate stimulus triggers) to prevent a competitor who aimed to punish specific behavioural patterns in each block. In this sense, our participants were reasons-responsive to the challenge represented by the current gaming environment. Voluntary actions elicited in this paradigm are taken to reflect how flexible or how adaptable reasons-guided voluntary action choices can be. By developing a series of environmental challenges and an analysis pipeline for quantifying adaptive autonomy for the first time, we found that people can become autonomous regarding both choice bias in favouring one action and also sequential biase in transition from one action to the next. However, participants had very limited ability to become free of reinforcement bias. We further showed, in a large sample, that the correlations between these three forms of adaptive

autonomy were minimal. This suggests distinct cognitive modules for these three forms of autonomy, rather than a common module or a single form. This finding contrasts with the view that a single cognitive control process underpins voluntary actions (Botvinick et al., 2001). We also demonstrated that model-based behaviour leads people to become more autonomous.

Measures of autonomy

Traditional experimental psychology struggles to investigate volition and autonomy, even though these are distinctive features of the healthy adult mind (Locke, 1690/1975). Experimental approaches to the study of volition often involved giving people a paradoxical instruction regarding 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 explicitly aimed to investigate human autonomous behaviour typically used competitive game contexts (Forder & Dyson, 2016; Wang et al., 2014; Wong et al., 2021), and have rarely considered subtypes of autonomy. We conceptualized three forms of autonomy, as freedom from three cognitively distinguishable types of action-generation strategy: *choice bias*, *sequential bias* and *reinforcement bias*. We attempted to evoke autonomous behaviour by programming a competitor to punish any lack of each type of autonomy. Using a statistical distance measure derived from information geometry, we quantified the extent to which people could break their own stereotypical choice pattern when punished. We found no strong correlations between these three types of adaptive autonomy, making a single common underpinning cognitive control function unlikely, and supporting the idea that behavioural autonomy involves dissociable components. 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 domain-specific top-down control independently regulates each particular form of autonomous behaviour: there are multiple ways to act freely, and it is important to consider *from what* an agent may be free. We studied choice biases, sequential biases and reinforcement biases, but other biases of free action doubtless also exist. We showed that, for example, an agent who becomes increasingly free from choice bias may yet be unable to free themselves from the biasing effects of reinforcement.

Relevance to classical neuropsychological tasks

Our study evokes behavioural 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). Next, the sequential bias we measured reflects executive control and working memory which are often assessed using random number generation tasks (Baddeley et al., 1998; Jahanshahi et al., 2000). In these tasks, people perform poorly at randomising numbers, tending to rely on inappropriate heuristic rules, and misunderstanding a concept of randomness (Baddeley et al., 1998; Ginsburg & Karpiuk, 1994). In contrast, the capacity to avoid the reinforcement bias is related to voluntary override of reward-seeking behaviour (Bechara et al., 1994; Lejuez et al., 2002) and to the balance between exploitation and exploration (Cohen et al., 2007). These classical executive functions are often seen as robust individual traits affecting any relevant task that draws on executive capacity (Neiman & Loewenstein, 2011; Ota et al., 2016; Ota et al., 2019). Here we instead see these stereotyped action selections as a state function of cognitive control processes. We developed a competitive game in which participants would need to avoid particular action preferences, action generating rules, or outcome dependences. Our results demonstrate that people can balance choice frequencies and break transitions between actions. Our results suggest that, when circumstances require, people can appropriately increase their behavioural autonomy.

In contrast, we found people could not avoid reinforcement bias. Neither positive reinforcement bias nor negative reinforcement bias adapted 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 regarding changing lose-shift behaviour than win-stay behaviour when adapting to new game rules (Forder & Dyson, 2016; Sundvall & Dyson, 2022). The experience of a negative outcome modulates the speed-accuracy trade-off on the subsequent trial (Desender et al., 2021; Dyson et al., 2018). Individual differences are also 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 failure 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. Rather, their choice strategy depended on how they interacted with the environment. Reinforcement learning updates the best action by a reward prediction error, while 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. Unlike classical reinforcement tasks, there was no explicit rewarding option that participants can exploit, because our task was designed to produce stimulus-independence and because participants choose along a continuum of action times (Libet et al., 1983), rather than choosing between discrete response categories. Nevertheless, both reinforcement learning and belief learning provided better fits to participants' performance than random selection. 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 intend and do (Amodio & Frith, 2006; Hampton et al., 2008).

We found that, when the competitor started predicting people's upcoming action, people also started predicting the competitor's action, switching their action-generating strategy away from reinforcement learning towards belief learning. This result suggests that people implicitly represented a model of the competitor's strategy (or the likelihoods of the competitor's actions) based on a sampling of previous trials. Volitional action choices were made not simply on the basis of what was rewarded on the previous trial, but plausibly because participants knew that the current choice would be better, based on their model of the competitor.

The shift from model-free reinforcement learning towards model-based behaviour was indeed a key for achieving greater autonomy. Using structural equation modelling, we demonstrated that a greater adaptive autonomy score was found in individuals who showed a greater shift to belief learning. While habitual behaviour is considered an action under control of model-free reinforcement learning (Dolan & Dayan, 2013), this finding clarifies a link between breaking habit and moving from model-free towards model-based behaviour. In our task, this process must have been implicit, because participants were not told anything about the algorithm the competitor used. 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.

Limitations

Our study has a number of limitations. First, one might question on theoretical grounds whether our paradigm offers a good testbed for volition. Somewhat similar competitive games have, after all, been proposed in studies of social cognition, decision-making and learning – without reference to volition. This criticism may be difficult to resist given that the precise definition of volition remains controversial. However, using a recent list of the proposed features of volition (Haggard, 2019), we have shown that the present task elicits actions that are not only stimulus-independent and non-habitual, but also goal-directed, thus satisfying classical philosophical criteria for volition. Further, our task involves motivated selection of when to act, similar to a classical neuroscientific study of volition (Libet et al., 1983).

Next, our participants completed a structured series of games against progressively sophisticated competitors in a fixed order. This design was logical given the hierarchical dependency of the different biases that we hypothesised could constrain the capacity for autonomous volitional action. For example, we reasoned that sequential biases of action choice (block 2) could only be studied after biases in action choice per se (block 1) were first controlled for. However, fixed order designs are subject to confounding effects of learning, fatigue and other time-dependent factors. For example, our participants' inability to adapt to reinforcement biases (block 3) may be contaminated by an element of fatigue. While such confounds cannot be excluded absolutely by our design, our findings of successful adaptation initially (blocks 1 & 2) followed by less successful adaptation (block 3), rule out progressive effects of a single factor such as fatigue.

We in effect examined this concern of the fixed order design in another pre-registered study (https://aspredicted.org/4u7y3.pdf). In this study, we introduced the condition where a virtual competitor merely behaves randomly and does not punish any choice patterns. We confirmed that the control group under this random, no-punishment condition did not show any behavioural changes whilst the experimental group reduced their choice bias scores when punished. Therefore, the presence or absence of behavioural changes we found is not simply explained by time-dependent factors.

We also note that the AICc difference between the hybrid learning rule and the stochastic selection model became relatively small at the punishment of action sequence (block 2), relative to the punishment of choice preference (block 1; Figure 5B). Some proportion of the participants (37.5% in Figure 5C) was not fully explained by the hybrid learning rule in block 2. That is being said, our model captured the participant's performance at

least at some extent. Models that better capture a transition between actions could be addressed in future.

Empiricist view versus nativist views of human autonomy

Our work is broadly compatible with an empiricist view of human autonomy 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. Such 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 decision bias 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 strong nativist view would suggest that autonomy is a state that occurs inside an individual’s mind and is independent of the external environment. However, our results imply that being sensitive to the contingencies of the external environment is key to autonomy, and that autonomous agents are characterised by an appropriate reaction to the restrictions that the environment places on their own actions. In this sense, autonomy can be seen a reasoned, goal-oriented response that occurs through interaction with an environmental context.

Conclusions

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 internal constraints of habitual responding, when a particular choice pattern or stereotypical behaviour 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 the elements of their biased action selection pattern, including repetition of choice, rule-based sequential action and actions of being guided by reinforcement. These appear to be three distinct forms of adaptive autonomy, rather than a single common cognitive mechanism for avoiding all decision biases, for example by strategic randomness. We show that becoming free from the effects of reinforcement is particularly difficult. By demonstrating that belief learning plays an important part in boosting autonomy, we argue for a strong connection between autonomous action and mentalising abilities. Finally, we have looked at behavioural adaptation and cognitive flexibility

of action choices through the lens of voluntary action. Skinner, in his *Beyond Freedom and Dignity* (Skinner, 1971/2002) argued that reinforcement-based guidance of conditioned responses obviated the need any cognitive construct of volition. We have argued a different position. At least in the context of competitive games, the key cognitive features of volitional action are precisely what allows flexible, adaptive and successful performance.

Competing interests

The authors declare no competing interests.

Author's contributions

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

Acknowledgments

This publication was made possible through the support of a joint grant from the John Templeton Foundation and the Fetzer Institute. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the Fetzer Institute. We are grateful to Mr Alexander Platt, Dr Annika Boldt, Dr Stefano Palminteri for their helpful review on the manuscript.

Data and code availability

The data frame for a representative participant is available in the supplementary data. The data frames for all the participants and the codes used to generate the figures will be posted.

References

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

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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001, Jul). Conflict monitoring and cognitive control. *Psychol Rev, 108*(3), 624-652. https://doi.org/10.1037/0033-295x.108.3.624

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

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

Desender, K., Ridderinkhof, K. R., & Murphy, P. R. (2021, Aug 20). Understanding neural signals of post-decisional performance monitoring: An integrative review. *Elife, 10*. https://doi.org/10.7554/eLife.67556

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

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

Ginsburg, N., & Karpiuk, P. (1994). Random Generation: Analysis of the Responses. *Perceptual and Motor Skills, 79*(3), 1059-1067. https://doi.org/10.2466/pms.1994.79.3.1059

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

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

Locke, J. (1690/1975). An essay concerning human understanding. *New York : Oxford University Press*.

Maoz, U., & Sinnott-Armstrong, W. (2022). Free Will: Philosophers and Neuroscientists in Conversation. *Oxford University Press*.

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

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

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

Roskies, A. L. (2010). How does neuroscience affect our conception of volition? *Annu Rev Neurosci, 33*, 109-130. https://doi.org/10.1146/annurev-neuro-060909-153151

Shadlen, M. N., & Roskies, A. L. (2012). The neurobiology of decision-making and responsibility: reconciling mechanism and mindedness. *Front Neurosci, 6*, 56. https://doi.org/10.3389/fnins.2012.00056

Skinner, B. F. (1971/2002). Beyond freedom and dignity. *Hackett Publishing*.

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

Weinberg, D. (2020). Freedom and addiction in four discursive registers: A comparative historical study of values in addiction science. *History of the Human Sciences, 34*(3-4), 25-48. https://doi.org/10.1177/0952695120945962

Weinberg, D. (2022, Sep). The particularity of freedom and recovery from addiction in practice: On the therapeutic limitations of generalising addiction science. *Int J Drug Policy, 107*, 103609. https://doi.org/10.1016/j.drugpo.2022.103609

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

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