Contrôle, agentivité et apprentissage par renforcement

Héloïse Théro

▶ To cite this version:

Héloïse Théro. Contrôle, agentivité et apprentissage par renforcement. Neurosciences [q-bio.NC]. Université Paris sciences et lettres, 2018. Français. NNT : 2018PSLEE028 . tel-02109235

HAL Id: tel-02109235
https://theses.hal.science/tel-02109235

Submitted on 24 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'École Normale Supérieure

Control, agency and reinforcement learning in human decision-making
Contrôle, agentivité et apprentissage par renforcement

École doctorale n°158

CERVEAU, COGNITION, COMPORTEMENT

Spécialité NEUROSCIENCES COGNITIVES

COMPOSITION DU JURY :

Mathias PESSIGLIONE, Président
Institut du Cerveau et de la Moelle épinière

Benjamin SCHEIBEHENNE, Rapporteur
University of Geneva

Markus ULLSPERGER, Rapporteur
Otto von Guericke Universität Magdeburg

Pierre-Yves OUDEYER, Membre du jury
Inria Bordeaux Sud-Ouest

Etienne KOECHLIN, Directeur de thèse
Ecole normale supérieure

Soutenue par **Héloïse THÉRO**
le 26 septembre 2018

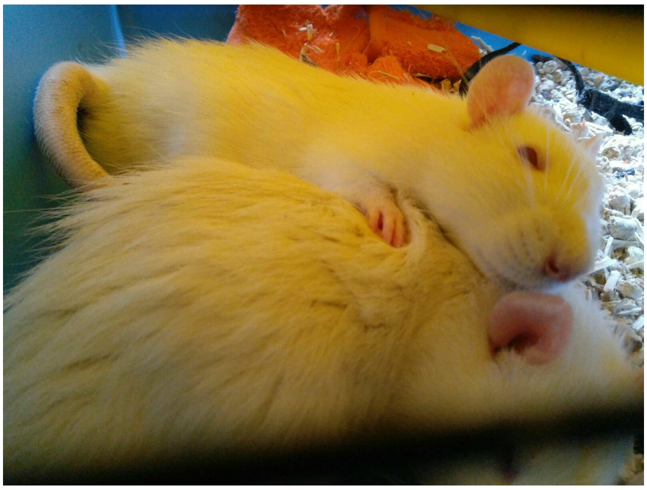
Dirigée par **Étienne KOECHLIN**

ENS | PSL

"Toute découverte, qu'elle soit philosophique, scientifique ou autre ne peut être considérée que comme un stade dans l'Histoire de l'homme et non comme une découverte de la *vérité. L'homme de la Terre a encore besoin de certaines croyances, de certaines théories, même si celles-ci les plongent dans l'erreur pour quelque temps. Une erreur correspond à un degré dans la quête de la vérité."*

Daniel Meurois et Anne Givaudan

À la mémoire de Nuage et Griffon



Contents

Chapter 1

General Introduction

1.1 Conditioning and reinforcement learning

We will first introduce how a very influential reinforcement learning algorithm allows us to understand many aspects of animal behavior and neural activity. At the turn of the twentieth century, a new approach called ’behaviorism’ tried to emulate physics by explaining animal behavior in terms of mechanics that could be easily measured. At first behaviorism fortified the old Cartesian wall between humans and animals: humans were thinking organisms who shape their environment; animals were mindless brutes whose behavior is conditioned by the environment. Then Burrhus Skinner tried to tear down this wall by getting rid of the mind. He claimed that the human learning process is no different from conditioning in animals, and that it could be described mechanically without resorting to nebulous terms like ’thought’ or ’consciousness’ (Fouts and Mills, 1997).

1.1.1 The notion of reward

Before the 1950s, the prevailing view held that the basic motivations, such as pain, pleasure and so on, probably involved excitation or activity of the whole brain. In 1953, Olds and Miller implemented by accident an electrode in a nerve pathway from the rhinencephalon. They observed the implemented rat learned to return to the portion of its environment where it had been given the electrical stimulation (Olds, 1956).

This demonstration of a learned place preference suggested that these stimulations were rewarding. They thus placed the animals in a box in which they could stimulate themselves by pressing the lever. The rats were then self-stimulating about once every five seconds. When they turned off the current (so that the animal’s pressing of the lever could no longer stimulate the brain), the animals kept pressing it only a few times before going to sleep (Olds and Milner, 1954).

Olds (1956) found that the strongest reward, or pleasure, came from stimulating the hypothalamus and certain mid-brain nuclei, hence describing the reward system for the first time. Later these brain areas were identified as receiving dopamine from the ventral tegmental area and substantia nigra (Schultz, Dayan, and Montague, 1997). We will later see why this neurotransmitter is important for reinforcement learning.

A pleasant stimulus is familiarly called a reward. But it should be noted that

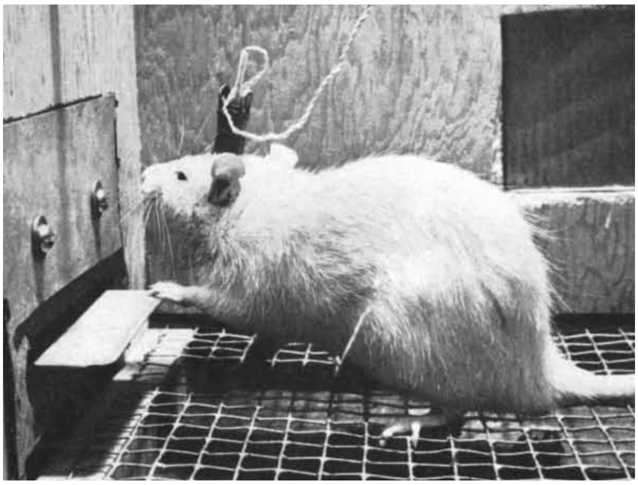


Figure 1.1: When the rat presses on the treadle, it triggers an electric stimulus to its brain, creating a self-stimulation circuit. Some of the animals have been seen to stimulate themselves for 24 hours without rest, and as often as 5,000 times an hour. (Figure reproduced from Olds, 1956.)

actual reward lies in active processes of the brain that reacts to a stimulus rather than the stimulus itself, as this experiment showed. A reward is actually a composite process containing several psychological components:

- *Liking*, which is the actual pleasure component or hedonic impact of a reward,
- *Wanting*, i.e., the motivation for reward, which makes the animal approach reward and avoid punishment,
- *Learning*, i.e., the associations, representations, and predictions about future rewards based on past experiences.

These different aspects are mediated by partly dissociable brain substrates. Within each reward component, there are further subdivisions and levels, including both conscious and non-conscious processing (Berridge and Kringelbach, 2008).

We have seen that the *learning* and *wanting* components of reward were present in the rats' electrical self-stimulation. But the challenge in the *liking* aspect is that it is very difficult to access such subjective 'pleasure' states in experimental work, particularly in animals. In humans, one can simply ask participants to verbally report or rate their subjective pleasure (O'Doherty, 2014). Humans implemented with 'pleasure' electrodes often displayed the same *wanting* behavior as the rats (Heath, 1972; Portenoy et al., 1986). But there was no clear evidence that electrodes caused real pleasure. A patient described "erotic sensations often intermixed with an undercurrent of anxiety. She also noted extreme thirst, drinking copiously during the session, and alternating generalized hot and cold sensations" (Portenoy et al., 1986).

Punishment is usually defined as the opposite of reward. A debate in cognitive neuroscience concerns whether the same brain areas, namely the ventral striatum and the ventromedial prefrontal cortex, represent reward as well as punishment (Bartra, McGuire, and Kable, 2013) or whether aversive value encoding and learning are organized in an opponent system, namely the insula and the dorsomedial prefrontal cortex (Garrison, Erdeniz, and Done, 2013). What is clear is that humans and

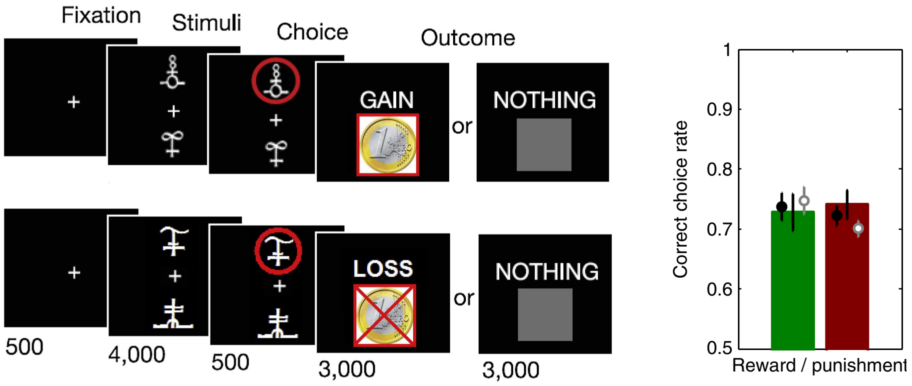


Figure 1.2: In each context (monetary gains or losses), options were associated with different outcome probabilities, so that the subjects' task was to learn which option was associated with either the highest reward, or the lowest punishment probability. Healthy subjects learnt similarly from reward and punishment. (Figure reproduced from Palminteri et al., 2012; Palminteri et al., 2015.)

animals are able to learn equally well by seeking rewards and by avoiding punishments (Pessiglione et al., 2006; Palminteri et al., 2015).

1.1.2 Classical and instrumental conditioning

Here we are mainly interested in the learning aspect of reward. Learning about stimuli or actions solely on the basis of the rewards or punishments associated with them is called associative learning or conditioning or reinforcement learning. Conditioning is traditionally separated into classical (or Pavlovian) conditioning, and

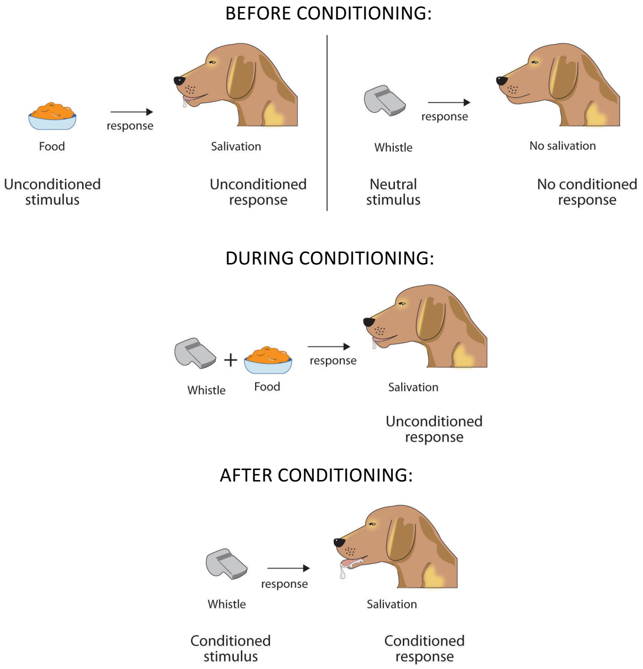


Figure 1.3: Before conditioning, the dog displays an 'unconditioned response': he salivates when food is put in his mouth. After repeatedly hearing a whistle before the arrival of food, the dog now salivates as soon as he heard the whistle, displaying a 'conditioned response'. (Figure adapted from www.savingstudentsmoney.org/psychimg.)

instrumental (or operant) conditioning. In Pavlovian conditioning, the rewards or punishments are delivered independently of any actions taken by the animal. Everyone knows Ivan Petrovich Pavlov was making dogs salivate with a bell, although not a lot of people understand why his discoveries were crucial for psychology.

Pavlov was originally interested in the physiology of digestion. Dogs are salivating as soon as food is put into their mouth (as we also do). Pavlov called this reflex ’unconditioned response’, as it is an automatic behavior that cannot be learned or changed. He discovered that an arbitrary signal, that could be a whistle, the vanilla smell or the view of a rotating object, can also cause salivation, if this arbitrary signal was repeatedly perceived just before the arrival of food. He called this learned behavior a ’conditioned response’. His results revealed that the most basic form of learning can be studied experimentally (Frith, 2013).

On the contrary, in instrumental conditioning, the actions of the animal determine what reinforcement is provided. As this PhD thesis focus on the link between control (i.e. how your actions can shape your environment) and reinforcement learning, instrumental conditioning is of particular interest for us.

By the time Pavlov was studying dogs, Edward Thorndike was putting a hungry cat into what he called ’puzzle box’, i.e., a box that could be opened if the animal pressed a lever or pulled a loop. He observed that cats were indeed able to learn to go out of the cage, but he wanted to understand how. He saw that cats could not learn by observation (i.e., by seeing another cat get out of the puzzle box), but only by trial-and-error (Frith, 2013). Thorndike called this associative learning the ’law of effect’, stating that “responses that produce a satisfying effect in a particular situation become more likely to occur again in that situation, and responses that produce a discomforting effect become less likely to occur again in that situation” (Thorndike, 1911).

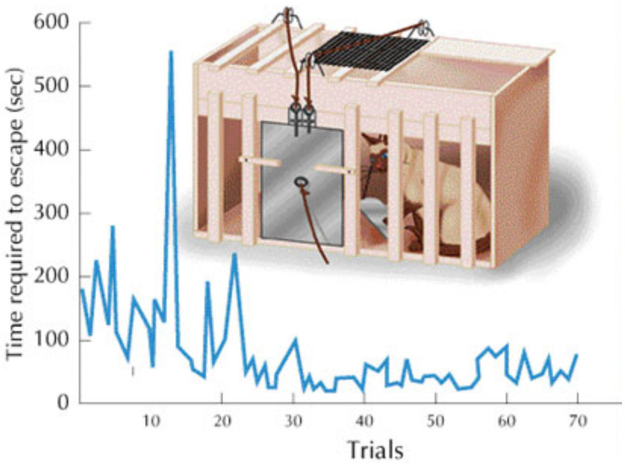


Figure 1.4: Thorndike placed cats in a puzzle box that could be opened if the cat pressed a lever. Thorndike noted that with each successive trial, it took the cat less and less time to escape on average. (Figure reproduced from commons.wikimedia.org.)

Skinner later generalized the use of boxes in which some actions are linked rewards or punishments, called ’Skinner boxes’. The self-stimulation box that we described earlier is a particular kind of Skinner box (Olds, 1956).

1.1.3 The TD(0) algorithm

The first evidence that animal learning can be described by a reinforcement learning algorithm came from a Pavlovian conditioning experiment. After conditioning, an animal's behavior indicates that the conditioned stimulus induces a prediction about the likely time and magnitude of the reward. Schultz, Dayan, and Montague (1997) deduced that no further learning should thus take place when the reward can be entirely predicted by the conditioned stimulus.

They recorded the activity of single dopamine neurons in alert monkeys while they were presented with stimuli and rewards. The majority of dopamine neurons (55 to 80%) are known to respond with short, phasic activations when animals touch a small morsel of apple or receive a small quantity of fruit juice to the mouth. Surprisingly, after repeated pairings of visual and auditory stimuli followed by reward, dopamine neurons change the time of their phasic activation from just after the time of reward delivery to the time of stimulus onset. In trials where the reward was not following the conditioned stimulus, dopamine neurons are depressed markedly below their basal firing rate exactly at the time that the reward should have been delivered.

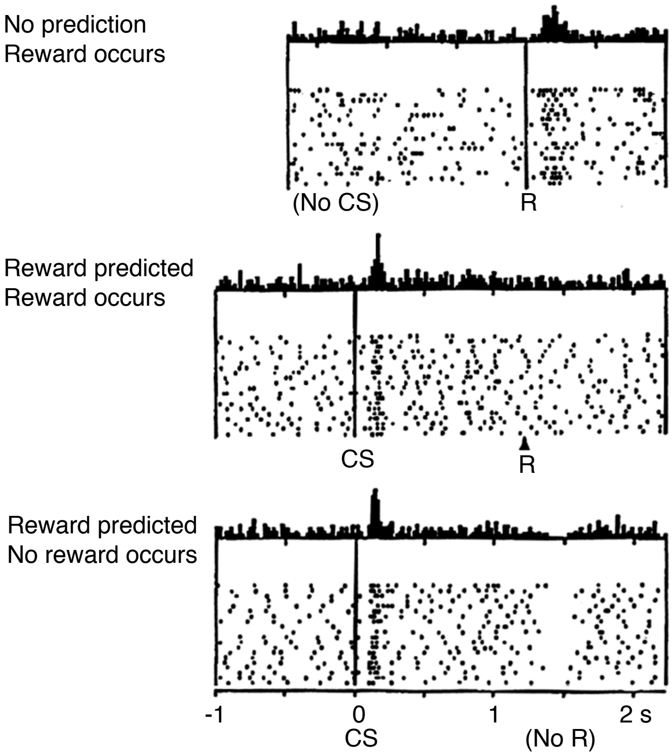


Figure 1.5: Dopamine as a reward error prediction. Each panel shows the peri-event time histogram (top) and raster of impulses (bottom) from the same monkey dopaminergic neuron (each line of dots shows one trial). Original sequence of trials is plotted from top to bottom. CS: conditioned, reward-predicting stimulus. R: primary reward. (Figure reproduced from Schultz, Dayan, and Montague, 1997.)

Dopamine outputs appeared to code for a deviation or error between the actual reward received and predictions of the time and magnitude of reward. The authors then paralleled this with the Temporal Difference error variable in the TD(0) algorithm, from the reinforcement learning framework:

$$\delta(t) = R(t) - V(t) \tag{1.1}$$

where $R(t)$ is the reward received on time t, and $V(t)$ the expected future reward associated with that stimuli or action on time t. The TD error $\delta(t)$ is used to improve the estimates of $V(t)$:

$$V(t+1) = V(t) + \alpha \times \delta(t) \tag{1.2}$$

where α is a learning rate parameter.

Reward prediction in both Pavlovian and instrumental conditioning tasks was later shown to rely on similar neural basis, namely the ventral striatum which receives the projections of dopaminergic neurons (O'Doherty et al., 2004). Therefore the TD(0) algorithm was used to explain instrumental, as well as Pavlovian, conditioning. Furthermore Pessiglione et al. (2006) investigated the effects of drugs enhancing or reducing dopaminergic function. They found that the magnitude of reward prediction error expressed in the striatum was indeed modified by dopamine treatments, and that participants treated with the dopamine enhancer better learned than participants treated with the dopamine blocker.

The reinforcement learning model was also used to explain two event-related potentials (ERPs) classically found in EEG measures: the error-related negativity (ERN) occurring after an erroneous response, and the feedback-related negativity (FRN) occurring after a negative feedback. Holroyd and Coles (2002) found these two types of ERPs to be generated when a negative prediction error signal is conveyed to the anterior cingulate cortex (ACC) via the mesencephalic dopamine system. The ACC appears to monitor errors, in order to then engage regulatory processes in the lateral prefrontal cortex to improve performance (Ridderinkhof et al., 2004).

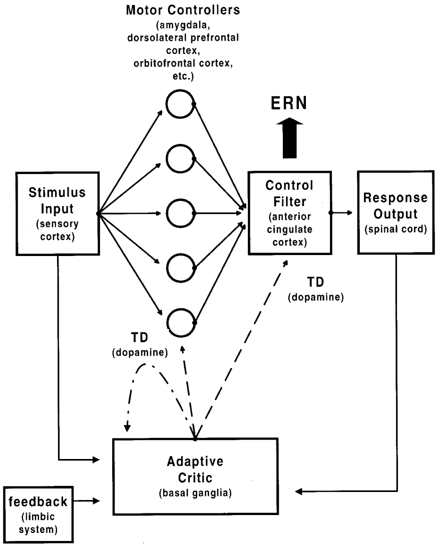

Motor Controllers
(amygdala, dorsolateral prefrontal cortex, orbitofrontal cortex, etc.)
ERN
Stimulus Input (sensory cortex)
Control Filter (anterior cingulate cortex)
Response Output (spinal cord)
TD (dopamine)
TD (dopamine)
Adaptive Critic (basal ganglia)
feedback (limbic system)

Figure 1.6: The reinforcement learning theory of error monitoring is linking the temporal difference error to the error-related negativity. (Figure reproduced from Holroyd and Coles, 2002.)

The remaining mystery in instrumental conditioning was to understand how actions can be selected on the basis of their corresponding reward prediction. Reinforcement learning models can use different action selection rules:

- the *'hardmax'* rule: always choosing the optimal action, i.e. the action associated with the highest expected reward,
- the *'softmax'* rule: choosing the actions probabilistically on the basis of the actions' relative expected reward,
- the *'ϵ-greedy'* rule: choosing the optimal action most of the time, but occasionally (with probability $1 - \epsilon$) substituting a random action.

Daw et al. (2006) compared the fit of models embodying these different action selection strategies. They found their participants' behavior to be better described by the softmax rule, with the probability of choosing action i taking the form:

$$P_i = \frac{e^{\beta \times V_i}}{\sum_j e^{\beta \times V_j}} \tag{1.3}$$

This particular reinforcement learning model instantiation (a TD(0) learning rule to learn the value predictions with a softmax rule for action selection) is now widely used to explain participants' choices in conditioning tasks (Gläscher and O'Doherty, 2010).

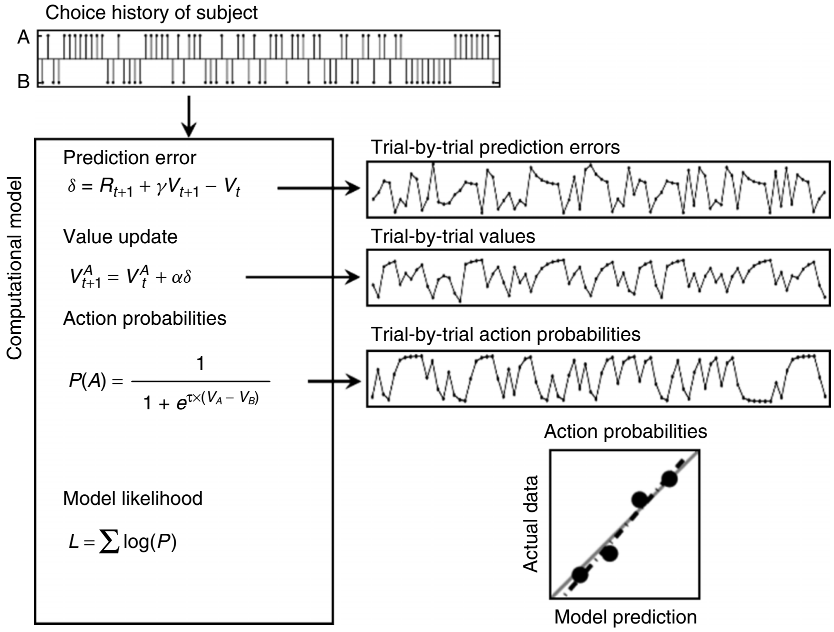

Choice history of subject
A
B
Computational model
Prediction error
$\delta = R_{t+1} + \gamma V_{t+1} - V_t$
Value update
$V^A_{t+1} = V^A_t + \alpha\delta$
Action probabilities
$P(A) = \frac{1}{1 + e^{\tau\times(V_A - V_B)}}$
Model likelihood
$L = \sum \log(P)$
Trial-by-trial prediction errors
Trial-by-trial values
Trial-by-trial action probabilities
Action probabilities
Actual data
Model prediction

Figure 1.7: A summary of the reinforcement learning model used for human cognition. (Figure reproduced from Gläscher and O'Doherty, 2010.)

However humans and many other animals spontaneously explore their environments, even when they are not under direct pressure for finding extrinsic rewards like food. Interestingly, curiosity-driven learning enables organisms to make discoveries to solve complex problems with rare or deceptive rewards (Oudeyer, 2018). Colas, Sigaud, and Oudeyer (2018) have recently added a Goal Exploration Process, which benefit from more focus on exploration, within a standard deep reinforcement learning algorithm. The addition of an explorative early phase improved the standard algorithm in challenging environments.

1.1.4 Computational optimality

In 1979, Sutton and Barto developed the idea of a 'hedonistic' learning system that *wants* something, that adapts its behavior in order to maximize a special signal from its environment. Reinforcement learning is now one of the most active areas in machine learning, with different subfields such as dynamic programming, temporal-difference learning, and function approximation (Sutton and Barto, 1998).

Machine learning studies the class of algorithms provided with a set of data and designed to 'learn-by-examples'. A typical distinction is made between supervised learning, in which data are assigned to their corresponding target, and unsupervised learning, in which the algorithm is simply provided with an unlabelled set of data. Reinforcement learning is often said to be minimally supervised because agents are not told explicitly what actions to take in particular situations, but must work this out themselves on the basis of the reinforcement they receive (Dayan and Abbott, 2001).

Formally, the goal of reinforcement learning is to learn of a behavioral strategy (a policy) which maximizes the long term sum of rewards (delayed reward) by a direct interaction (trial-and-error) with an unknown and uncertain environment. Finite Markov Decision Processes are a classical formalization of sequential decision making, where actions influence not just immediate rewards, but also subsequent situations, or states.

The learner and decision maker is called the *agent*. The thing it interacts with, comprising everything outside the agent, is called the *environment*. These interact at each of a sequence of discrete time steps: $t = 0, 1, 2...$ At each time step t, the agent preceives some representation of the environment's state $S_t \in S$, where S is the set of possible states, and on that basis the agent selects an action $A_t \in A(s_t)$, where $A(s_t)$ is the set of actions available in state S_t. One time step later, in part as a consequence of its action, the agent receives a numerical reward, $R_{t+1} \in R \subset \mathbb{R}$, and finds itself in a new state, S_{t+1} (Sutton and Barto, 2017).

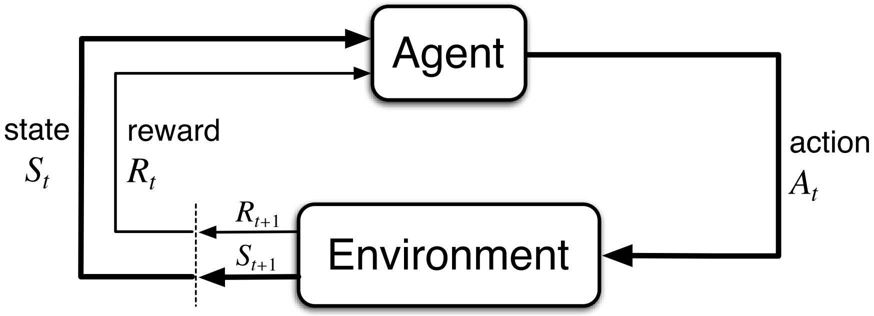


Figure 1.8: The agent-environment interaction in a Markov decision process. (Figure reproduced from Sutton and Barto, 2017.)

The use of a reward signal to formalize the idea of a goal is one of the most distinctive features of reinforcement learning. According to the discounting approach, the agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses A_t to maximize the expected *discounted return*:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{1.4}$$

where γ is a parameter, called the discount rate ($0 \leq \gamma \leq 1$).

Almost all reinforcement learning algorithms involve estimating value functions, that estimate how good it is for the agent to be in a given state, with respect to particular ways of acting, called policies. Formally the state-value function for policy π, denoted $v_\pi(s)$, is the expected return when starting in s and following π thereafter:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|S_t = s] \tag{1.5}$$

If the agent is following policy π at time t, then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$. The Bellman equation for v_π expresses a relationship between the value of a state and the value of its successor states:

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')] \tag{1.6}$$

for all $s \in S$.

The optimal policy to follow is defined as having the optimal state-value function, denoted v_*:

$$v_*(s) = \max_\pi v_\pi(s) \tag{1.7}$$

for all $s \in S$.

Interestingly the TD(0) algorithm has been shown to slowly converge to the optimal solution, given some conditions as a Markovian environment and an action selection rule allowing for some exploration (as the softmax or the ϵ-greedy rules).

We have seen how the reinforcement learning framework can explain learning by associations driven purely from reward and punishment. Now we will further develop why the notion of control is important to take into account in reinforcement learning.

1.2 Conditioning and instrumental control

Here we will review how control and reinforcement learning are two interdependent concepts, and how manipulating instrumental control in conditioning experiments have led to rich discoveries.

1.2.1 Superstitious behaviour

According to the reinforcement learning theory, learning will take place as long as an action and a subsequent reward are contingent, and not necessarily when the action is actually *causing* the reward. Skinner (1948) therefore hypothesized that conditioning protocols could create superstitious behavior, if rewards were to be given in a random manner. Eight hungry pigeons were put in an experimental cage with a food hopper. Food was given to the pigeons at regular intervals, with no reference whatsoever to the bird's behavior. After a brief delay, six pigeons exhibited a clear and repetitive behavior between the food arrivals: turning counter-clock wise two or three times, pecking or brushing movements towards the floor, hopping from right to left, etc.

Skinner interpreted the results as such: "The experiment might be said to demonstrate a sort of superstition. The bird behaves as if there were a causal relation between its behavior and the presentation of food, although such a relation is lacking. There are many analogies in human behavior." (Skinner, 1948).

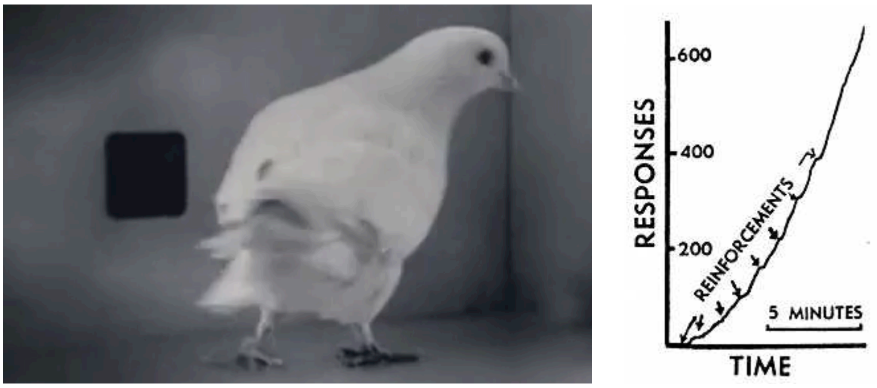


Figure 1.9: The response of hopping from right to left have been observed and mechanically recorded in a pigeon placed in a cage with rewards given at regular intervals. The arrows indicate the automatic presentation of food at one-minute intervals without reference to the pigeon's behavior. The bird does not respond immediately after eating, but when 10 or 15 or even 20 seconds have elapsed, it begins to respond rapidly and continues until the reinforcement is received. (Figure reproduced from Skinner, 1948.)

Superstitious, magical, and pseudoscientific thinking refer to ungrounded beliefs that are not supported by current evidence (Lindeman and Svedholm, 2012). Such beliefs are indeed widespread in people: two in five Europeans are superstitious (European Commission, 2010) and three-quarters of the American population believes in the authenticity of one or more paranormal processes (Moore, 2005). Interestingly, reinforcement learning can be used to explain complex behaviors such as superstitious actions.

1.2.2 Learned helplessness

Reinforcement learning is often of little help to understand moods or emotions (but see Eldar and Niv, 2015 for a counter-example). Resignation is a typical marker of depression and is highly correlated to neuroticism, a fundamental personality trait characterizing a persisting tendency to experience negative emotions (Jeronimus et al., 2016). Seligman, an American psychologist, identified resignation as a mood that can be learnt, instead of being a fixed individual trait. The theory of 'learned helplessness' postulates that, as humans are naturally prone to avoid suffering, resignation has to be learnt by repeatedly experiencing negative events, over which we have no control.

This hypothesis was exposed in a seminal study conditioning dogs with electrical shocks. Seligman and Maier (1967) used three groups of dogs. The first was a control group, placed in a hammock without any treatment. The second group was exposed to electrical shocks that could be suppressed when the dogs pressed a panel with their nose, and the dogs indeed were able learn this association. The third group was also exposed to electrical shocks, but without the possibility to alleviate them. Group 3 dogs were yoked to Group 2 dogs, so that both groups would receive the same average duration of shock.

The next day the animals were installed in a different environment, a shuttlebox escape, from which they could escape by jumping a partition. When again exposed to electrical shocks, 90% of the dogs in both Groups 1 and 2 learned to escape by crossing the barrier, while two third of the Group 3 dogs laid down passively during the shocks, failing to escape the shuttlebox.

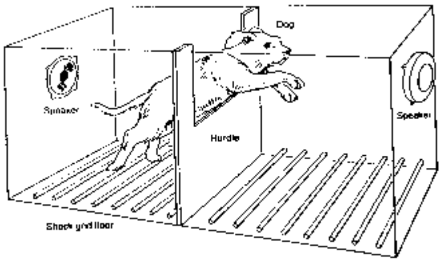


Figure 1.10: After receiving controllable or uncontrollable electrical shocks, the dogs in Seligman and Maier (1967)'s experiment were placed in a compartment from which they could easily escape. When again exposed to electrical shocks, animals having received controllable shocks easily found the solution to escape (as in the figure) while dogs previously subjected to uncontrollable pain laid motionless. (Figure replicated from Swenson's online lecture.)

Although this phenomenon was already known at the time (e.g. Overmier and Leaf, 1965), it was mainly interpreted as an interference effect (Adams and Lewis, 1962): animals were thought to fail escaping the shuttlebox because they had learnt in the first part of the experiment some behavior that interfered with the escaping behavior necessary in the second part. Seligman was the first to postulate and prove that this lack of initiative subsequent to uncontrollable shocks comes from the lack of control experienced. In a crucial experiment, any possible interfering behavior was prevented by immobilizing Group 3 dogs with curare while exposing them to uncontrollable shocks. Later, when placed in the compartments where electrical shocks were given, Group 3 dogs were again immobile, although they could not have developed an interfering behavior from their previous shock exposure (Overmier and Seligman, 1967).

Another interpretation was proposed to explain the helplessness behavior observed. From a behaviorist point of view, animals only learn an association between a specific response and an event, and they are unable to learn that they lack control over events. Thus it was claimed that in Seligman and Maier (1967)'s experiment, when the animals first receive unescapable shocks in the hammock, shock offset was occasionally paired with not moving. This could reinforce the association of not moving with shock offset, explaining the absence of movement when animals were later placed in the escapable shuttlebox. As Skinner (1992) has shown, non-existent action-outcome relationships can indeed be conditioned in a situation of lack of control.

But Maier (1970) refuted this behaviorist-compatible interpretation of learned helplessness by testing a variant of the original task in which Group 2 dogs could suppress shock by holding still, and not by pressing a lever. Similarly to their previous results, they found that dogs subsequently succeeded to learn to escape the shuttlebox by jumping the partition. These results grounded the theory of learned helplessness: experiencing uncontrollable negative events leads to a general helplessness state of mind. This experiment crucially showed that beyond learning pure response-outcome associations, animals are able to learn a general lack of control. That finding, among other results, constituted the beginning of a cognitive theory of animal behavior.

Almost exactly at the same time as Seligman and Maier (1967), the same paradigm was used on rats, but this time to study the physiological effects of escapable and unescapable electrical shocks (Weiss, 1968). Triplets of rats were placed into a restraining and escape-avoidance cage. One rat was randomly assigned to the Nonshock group, and its tail electrode was disconnected to the electric generator. One rat was assigned to the Avoidance group and the electrical shocks received could be terminated when the animal touched the copper plate in front of him with its nose. The rat assigned to the Yoked group received the same electrical shocks as the Avoidance rat, and the touch plate in front of it was not connected to the electrical shock relay.

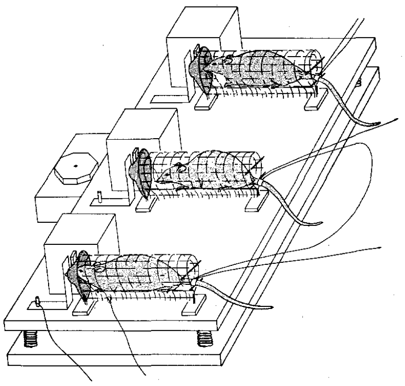


Figure 1.11: An illustration of a triplet of rats used in Weiss (1968)'s protocol. The lower rat is part of the Avoidance group, the middle rat of the Yoked group and the upper rat of the Nonshock group. (Figure reproduced from Weiss, 1968.)

Weiss (1968) found that Yoked rats showed a greater decrease in body weight and more extensive gastric lesions (also known as stress ulcers) than Avoidance rats. Thus the physiological effects of stress could be decreased by being able to perform a response that controlled the electrical shocks, that Weiss called a coping response. This study shows that exerting control over negative events is an important factor not only to predict how an animal will react to the next aversive events, but also to reduce stress and its noxious consequences.

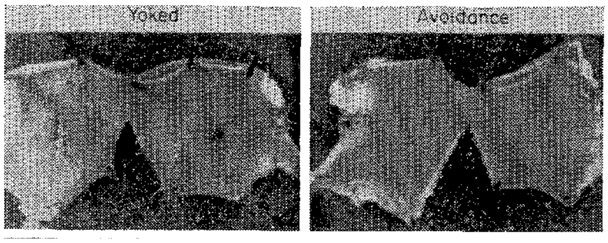


Figure 1.12: Photographs of the stomachs of one Yoked example rat and one Avoidance example rat from Weiss (1968)'s study. An independent judge, blind to the experimental condition, identified a lesion as a clearly visible defect or break in the mucosa, which was often accompanied by hemorrhage. She found more extensive gastric lesioning in the Yoked than in the Avoidance group. (Figure reproduced from Weiss, 1968.)

Learned helplessness was found to be characterized by a constellation of behavioral changes, that go well beyond a reduced escaping behavior or an increase in stress markers. Uncontrollable stressors also tended to reduce swimming when the animal was placed in water; reduce aggression and social dominance; produce neophobia, exaggerated fear and fear conditioning; reduce social interaction; produce opioid analgesia; reduce learning of instrumental responses for appetitive rewards; increase rewarding effects of opiates, and so on (see Maier and Watkins, 1998 for a review).

We have seen that instrumental control through avoidance options can produce substantial reductions in stress. But conditions were also found which produced more pathology in animals able to perform a coping response than in helpless animals. Weiss (1971) put rats in a situation in which they had to make a bar press response to avoid shock – but successful avoidance was signaled by an aversive blast of loud noise. In effect, the rats had to choose between two negative outcomes: shock versus noise, and they developed ulcers comparable to those of the helpless rats exposed to inescapable shock. Difficult decisions (as choosing the lesser of two evils) thus instigate costly inner processes.

Although learned helplessness is sometimes called 'behavioral depression' (Weiss et al., 1981), it should be noted that consequences of exposure to uncontrollable stressors are as similar to depressive symptoms as to those of extreme anxiety. Moreover, learned helplessness was found to be sensitive to both anti-depressants and anxiolytics (Maier and Watkins, 2005). This is perhaps not surprising since the development of both depression and anxiety may be influenced by stress, particularly uncontrollable stress. But caution should be used when extrapolating the learned helplessness literature to explain the development of psychiatric diseases as major depressive disorder.

1.2.3 The notion of instrumental contingency

The intuitive notion of helplessness entails the belief that no action that one does will matter, and therefore can be defined as a perceived lack of control. Maier and Seligman (1976) defined the notion of controllability in an early review of their work. They called $p(RF|R)$ the conditional probability of an outcome or reinforcer RF following a response R (at $p(RF|R) = 1$, every response produces a reinforcer; at $p(RF|R) = 0$, a response never produces a reinforcer). Important events can sometimes occur when no specific response has been made, and they called $p(RF|\overline{R})$ the conditional probability of a reinforcer RF following an absence of response $\overline{R}$.

They defined that a response stands in a relation of control to a reinforcer if and only if:

$$p(RF|R) \neq p(RF|\overline{R}) \tag{1.8}$$

And conversely a reinforcer was said to be uncontrollable if $p(RF|R) = p(RF|\overline{R})$ for all possible responses R.

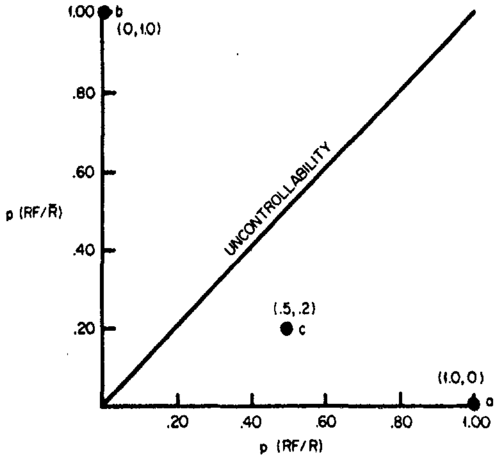


Figure 1.13: An outcome is defined as uncontrollable when the probability of the outcome or reinforcer RF following a response R is the same as the probability of the outcome RF following the absence of response $\overline{R}$. (Figure reproduced from Maier and Seligman, 1976.)

Independently from Maier and Seligman (1976), Hammond (1980) also defined instrumental contingency as the difference between probabilities of a reward R in presence or absence of an action A:

$$\Delta p = p(R|A) - p(R|\overline{A}) \tag{1.9}$$

If there is a causal action-reward relationship, the reward is said be 'contingent' with the action.

1.2.4 Instrumental contingency and behavior

In contrast with Skinner (1948)'s results, Hammond (1980) was the first to demonstrate that animals are sensitive to the causal relation between response and reward. Hungry rats were trained to press a lever under a schedule in which the first press in each one-second period is followed the delivery of a food pellet with a fixed probability. The causal relationship between lever pressing and food delivery was then degraded by increasing the probability that a food pellet will be delivered at the end of any second in which the animal does not press the lever. When the two probabilities of reward in presence and absence of lever pressing were set equal, pressing had no effect on the likelihood of the reward.

The important feature of this causal manipulation is that the contingency was degraded without altering the probability that a response was paired with a reward. Still, enhancing the probability of a reward in the absence of the response depressed the rats' instrumental performance.

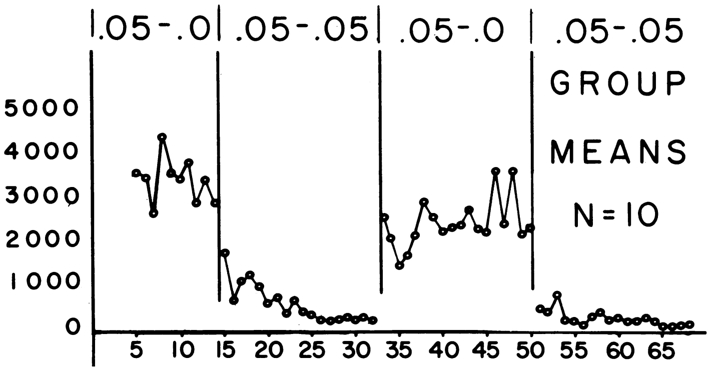


Figure 1.14: The mean response rate for ten rats accross different experimental sessions. The numbers above indicate the probability of a reward in the presence and absence of a response, thus .05-0 indicates an instrumental contingency, while .05-.05 indicates no contingency. (Figure reproduced from Hammond, 1980.)

Human participants were also found to be sensitive to the implemented contingency. Liljeholm et al. (2011) manipulated both the probability of a reward given an action $p(R|A)$, and given no action $p(R|\overline{A})$. They found that the participants' mean response rate was lower when contingency was degraded similarly to the rats' behavior in Hammond (1980)'s experiment.

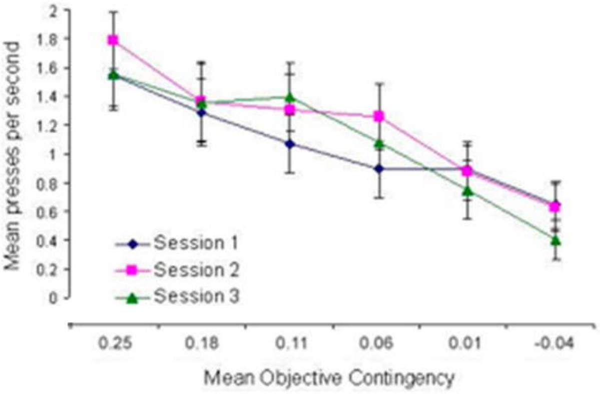


Figure 1.15: Mean presses per second in human participants across blocks sorted in descending order by objective contingency. (Figure reproduced from Liljeholm et al., 2011.)

Interestingly, Maier and Seligman (2016) drew a distinction between objective and subjective helplessness in a recent review. An animal is objectively helpless with respect to an outcome if this outcome is uncontrollable by any possible response. But being subjectively helpless is another matter. The animal must ′detect′ the lack of contingency as defined above and so must have expected that in the future the shock would be independent of its responses. Thus by experiencing an objective uncontrollability, an animal can develop a subjective helplessness. In the next section, we will see how the perception of control can differ from the actual control in human explicit reports.

1.3 The perception of control

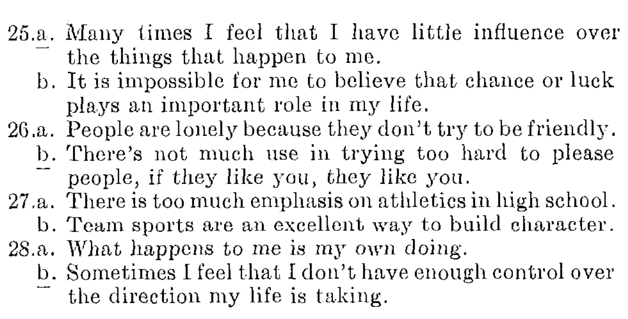
We will now review psychological studies on humans investigating the effects of control *perception* (and not control itself) on explicit reports and behavior.

1.3.1 The locus of control

The locus of control refers to people disposition to believe their fate either to be in their own hands or to be the consequence of external factors beyond their personal control. In an influential article, Rotter has hypothesized that the perceived locus of control would greatly influence how people will learn from reinforcements: "A person who is looking for an unusual brand of tobacco and is finally able to find it will return to the same place where he was reinforced before when he needs tobacco again. However, an individual who needs money and finds a five dollar bill in the street is not likely to return to that spot to look for a five dollar bill when he needs money." (Rotter, 1966).

An experiment was undertaken comparing verbal expectancies for future reinforcement under conditions of chance and skill learning. Phares (1957) used color matching as an ambiguous task and instructed half of the subjects that the task was so difficult as to be a matter of luck and the other half of his subjects that success was a matter of skill and that previous research had found some people to be very good at the task. For both conditions, a fixed order of partial reinforcement (right or wrong) was used. To measure the participants' expectancy, they were asked the number of chips they would bet on their probability of being correct on the succeeding trial. As Rotter has hypothesized, the increments and decrements of expectancy following respectively success and failure, were found to be significantly greater under skill instructions than under chance instructions.

Rotter developed and validated a questionnaire, the Internal-External Locus of Control Scale (I-E scale), to assess people's locus of control. Individuals who believe personal outcomes are contingent largely on their own behavior and attributes are said to have internal locus of control. On the other hand, people with external locus of control feel predominantly governed by other powerful individuals, institutions, luck, chance and so on. Rotter (1966) found that the scores on the locus of control questionnaire could explain individual differences in learning or not from reinforcers.



25.a. Many times I feel that I have little influence over the things that happen to me.
b. It is impossible for me to believe that chance or luck plays an important role in my life.
26.a. People are lonely because they don't try to be friendly.
b. There's not much use in trying too hard to please people, if they like you, they like you.
27.a. There is too much emphasis on athletics in high school.
b. Team sports are an excellent way to build character.
28.a. What happens to me is my own doing.
b. Sometimes I feel that I don't have enough control over the direction my life is taking.

Figure 1.16: Examples among the 29 items in the Internal-External Locus of Control Scale. The final score is given by the number of underlined items (item number 26 is a 'filler', as it is not related to locus of control). (Figure reproduced from Rotter, 1966.)

The I-E scale developed by Rotter has remained the most popular tool for measuring locus of control (Twenge, Zhang, and Im, 2004). A quantitative meta-analysis found an internal locus of control to predict many favorable work outcomes, such as positive task and social experiences, and greater job motivation (Ng, Sorensen, and Eby, 2006). Locus of control is an important part of a trait termed core self-evaluation, the other parts being self-esteem, self-efficacy, and emotional stability. Core self-evaluation was shown to be the best predictor for job performance and work and life satisfaction (Judge, 2009).

There is a general trend for paranormal beliefs to be associated with an external locus of control (Dag, 1999; Tobacyk, Nagot, and Miller, 1988). As a group, paranormal believers are inclined to feel specially vulnerable to external forces beyond their control (Irwin and Watt, 2007). This correlation suggests an interesting relationship between locus of control and the development of superstitious beliefs.

1.3.2 Illusions of control

In a seminal field study involving 631 adults, Langer (1975) showed that people tend to overestimate the probability of a positive outcome, and that this overestimation was based on factors that cannot rationally play a causal role in obtaining the outcome. For example, she found that response familiarity, or practice, on a chance task resulted in greater confidence in winning than when there was no practice. Increased confidence also resulted when the apparatus was controlled by the subject rather than the experimenter, even though in both instances the subject determined the response that would be made.

There are many naturalistic situations in which people fail to accurately judge a lack of contingency. Henslin (1967) studied dice playing and noted that dice players clearly behave as if they were controlling the outcome of the toss. They would throw the dice carefully and softly if they wanted low numbers, and to throw it hard for high numbers. They also believe that effort and concentration will pay off.

Blanco, Matute, and Vadillo (2011) investigated the effect of behavior on the illusion of control. Although they implemented no contingency between the participants behavior (a key pressing) and the outcome (a fictive patient recovering from a disease), subsequent judgements of contingency were positive, suggesting they developed an illusion of control.

Crucially, active participants (the participants that pressed the key the more often) were more prone to develop the illusion of control than those who responded less often. Blanco, Matute, and Vadillo (2011) suggested that this correlation could emerge from a cognitive dissonance phenomenon: the more participants have responded, the more prone they are to judge that their effort was not in vain.

In another experiment, Langer and Roth (1975) asked participants to guess the result of 30 coin toss. When feedback were manipulated to give participants an early, fairly consistent pattern of successes (although there was always in total 15 wins and 15 losses), participants predicted significantly more successes on future trials than those experiencing a random outcome sequence. In another experiment, the perception that one is causing a successful outcome was enhanced merely by the increased frequency of that outcome (Jenkins and Ward, 1965).

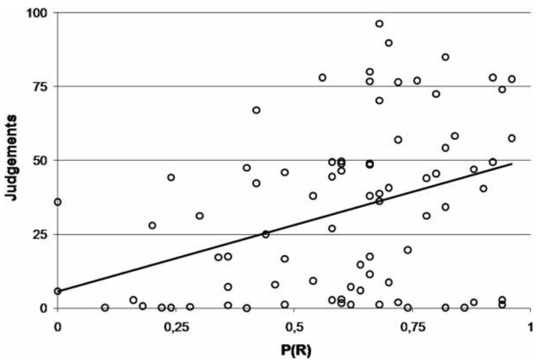


Figure 1.17: Contingency judgements as a function of probability of responding, P(R). (Figure reproduced from Blanco, Matute, and Vadillo, 2011.)

The illusion of control can thus be partly explained by a human bias to more attribute to oneself success than failure. Such link between perceived control and positive outcomes could explain why depressed individuals – who think less often of success – are not as likely as others to over-perceive control of successful outcomes (Alloy and Abramson, 1979).

1.3.3 An analytical perception of control

Illusions of control appear to be common in natural settings. But we also saw that humans and animals can be sensitive to instrumental contingency of their actions (Hammond, 1980; Liljeholm et al., 2011). Indeed, when humans are instructed before the beginning of the experiment to behave scientifically and to assess the response-outcome relationship, most studies found participants' judgments to be strong linear functions of the programmed contingencies (Chatlosh, Neunaber, and Wasserman, 1985; Wasserman, Chatlosh, and Neunaber, 1983).

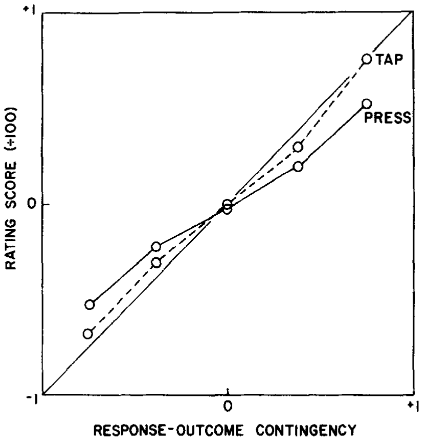


Figure 1.18: Scaled contingency ratings between a telegraph key operation and the illumination of a brief light, for each of the five levels of response-outcome contingency implemented. The 'Tap' group of participants was asked to produce a brief response by simply taping on the telegraph key, while the 'Press' group was asked to produce a continuous response by pressing the key for a variable length of time. (Figure reproduced from Wasserman, Chatlosh, and Neunaber, 1983.)

In another study using a free-operant contingency, participants could again press a key on a computer keyboard, which was associated with an outcome on the computer screen (Shanks and Dickinson, 1991) . But interestingly, participants in one group were asked to judge the effectiveness of the action in causing the outcome, while those in a second group were asked to maximize their points score under a payoff schedule. They observed low and constant proportions of responses as well as accurate judgments in the group instructed to assess control, and higher proportions of response in the group instructed to maximize outcomes, although they used a response cost to prevent the tendency to over-respond in the latter. They hypothesized that the natural performance strategy for maximizing reinforcement probably differs from the one to identify contingencies.

Trying to prove their hypothesis, Matute (1996) exposed participants to uncontrollable outcomes (as the termination of an aversive noise or a more neutral event as a beep). Half of the subjects was instructed to obtain the outcome (this condition was called 'naturalistic'), while the other half was instructed to respond on 50% of the trials and to assess their control over the outcome (corresponding to an 'analytic' condition). Subjects in the naturalistic condition tended to respond at almost every opportunity and developed a strong illusion of control. This illusion may simply be a collateral effect of a high tendency to respond, preventing them from learning that the outcome would have occurred with the same probability if they had not responded. By contrast, subjects in the analytic condition made accurate judgments of control.

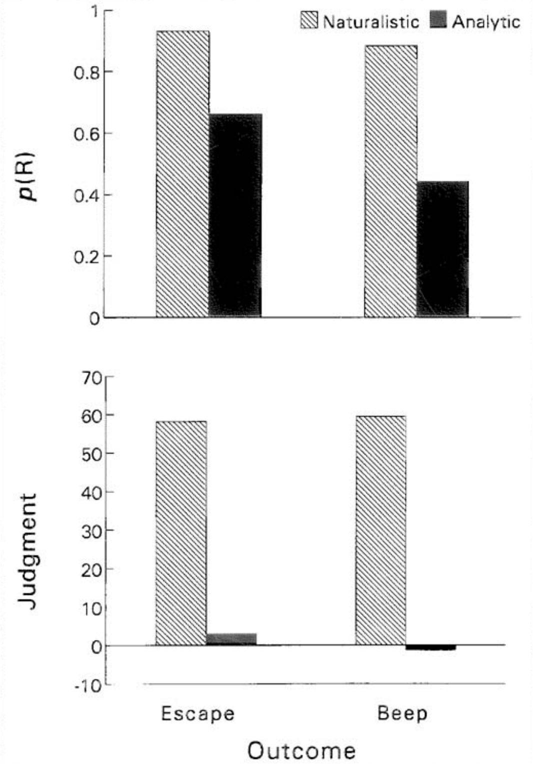


Figure 1.19: Mean probability of responding (p(R); top panel) and mean judgment of control (bottom panel) in naturalistic and analytic conditions. A judgement of 0 indicates an accurate perception of response-outcome independence. Escape represents the aversive noise condition, and beep the neutral stimuli condition. (Figure reproduced from Matute, 1996.)

1.3.4 Learning biases can explain illusions of control

We will now see how a computational perspective in sequential learning tasks can explain the emergence of illusions of control .

In a variety of behavioral tasks, subjects have been observed to readily alter their behavioral strategy in response to recent trends of stimulus statistics, even when such trends are spurious. Interestingly, this behavioral trend can be reproduced by an optimal Bayesian model under assumptions of statistical non-stationarity, while the same model under assumptions of stationarity would correctly infer an absence of control in a random environment (Yu and Cohen, 2009; Zhang, Huang, and Yu, 2014) . The participants' internal assumptions were then 'reverse-engineered': a random environment was perceived as actually changing about once every four trials (Yu and Cohen, 2009), although large inter-individual differences were found (Zhang, Huang, and Yu, 2014).

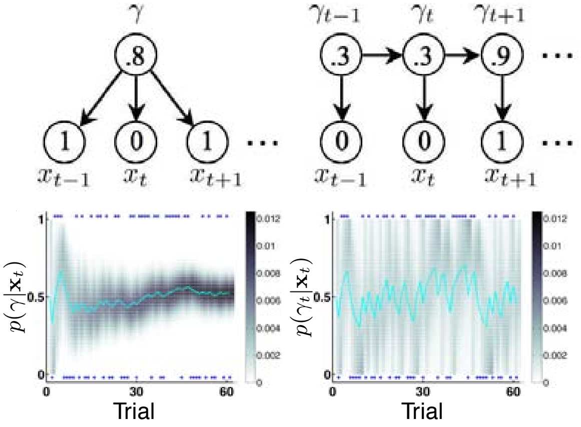


Figure 1.20: Graphical models of two Bayesian models assuming fixed and changing Bernoulli parameters (top panels). The two models made different inferences when observing the exact same sequence of truly random binary observations ($\gamma = .5$, bottom panels). (Figure reproduced from Yu and Cohen, 2009.)

Yu and Cohen (2009) conclude that it is very difficult to discriminate a truly randomized sequence, which by chance would contain runs of repetitions and alternations, from one that has changing biases for repetitions and alternations over time. Bialek (2005) also found plausible models allowing for changing biases to lead to surprisingly high probabilities of misidentifying random sequences as biased. Therefore if people assume they live in a changing environment, this belief will create an illusion of control, even in a random environment.

Recently, Nassar et al. (2010) have used a novel task to characterize how human subjects adapt their behavior in a changing task. They found that most subjects behaved as if they substantially overestimated the implemented volatility, suggesting that people tend to assume their environment is more changing than it really is. This misperception of volatility can be the source of people frequent illusions of control.

Lefebvre et al. (2017) have simulated the TD(0) algorithm that we previously described, in a task with a null instrumental contingency (as the reward probabilities were symmetric: 25% and 25% for both possible actions). This model and their participants were found to display transient preferred responses, as the Yu and Cohen (2009) Bayesian model with an assumption for volatility did.

Furthermore, some of Lefebvre et al. (2017) participants were better explained by a TD(0) model with different learning rates associated with positive and negative outcomes. These participants displayed a pronounced preference for one option (although both options were equally rewarding), as did the model with a higher positive than negative learning rate. Therefore, illusions of control can also be strengthened by an asymmetric update of positive vs. negative outcomes.

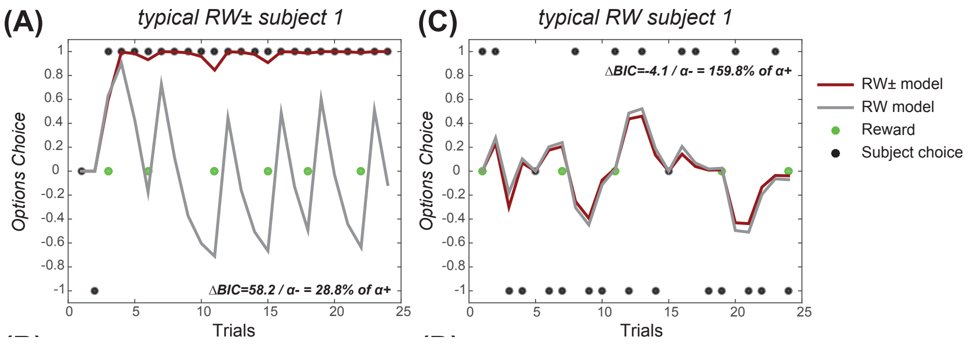


Figure 1.21: Behavioral choices and model simulations of a typical RW± participant (whose behavior is best fitted by a model with different positive and negative learning rates, left panel) and RW participant (whose behavior is best fitted by a model with a unique learning rates, right panel). (Figure reproduced from Lefebvre et al., 2017.)

A model has also been developed to explain transient, instead of global, illusions of control, as people often take previous outcomes into account when making predictions about random events. The prediction that the next outcome will be different from the previous one is often referred to as expectation of negative recency (for example when roulette players bet on red after the wheel has just landed on black). Scheibehenne and Studer (2014) analyses on a student sample revealed that prediction strategies varied across participants. Importantly, the different expectation patterns could be accounted for by a drift model that considers how often the same event has previously occurred in a row.

This section focused on the different biases arising when asking participants to report their perception of outcome control, and we have seen how computational models with certain assumptions can explain them. In the next section, we will focus on the perception of *action* control, often called sense of agency or subjective control.

1.4 The sense of agency

Sense of agency can be defined as the feeling that we control our actions, and through them effects in the outside world (Haggard and Chambon, 2012). Other researchers prefer to talk about the experience of conscious will (Wegner, 2003), following Hume's proposition to define will as a feeling (Hume, 1739).

1.4.1 The free will controversy

A crucial question is when does conscious will appear in the events surrounding actions? Kornhuber and Deecke (1965) have asked participants to move their right index finger at some arbitrary time in the following few seconds. Continuous recordings were made of electrical potential at several scalp electrodes while the actual time at which the finger moved was precisely measured by electromyography. Brain electrical activity was found to start increasing about 800 milliseconds before the voluntary finger movement, in the left and right precentral and the midparietal regions. This activity was called the readiness potential. It appeared that a general bilateral readiness for voluntary action later resolved into a more localized activation of the area responsible for the specific action, peaking 50 ms before the action unfolds.

Then the following question is when exactly in this sequence the person experience conscious will. Libet (1985) also asked participants to move their finger spontaneously, but this time while they were watching a clock. A spot of light was revolving each 2.56 seconds in a clockwise path around the circumference of the screen. The participant's task was simply to report for each finger movement where the dot was on the clock when he experienced "conscious awareness of wanting to perform a given self-initiated movement". The conscious willing of finger movement occurred at a significant interval *after* the onset of the readiness potential, but also at a significant interval before the actual finger movement (and also before the actual awareness of the movement).

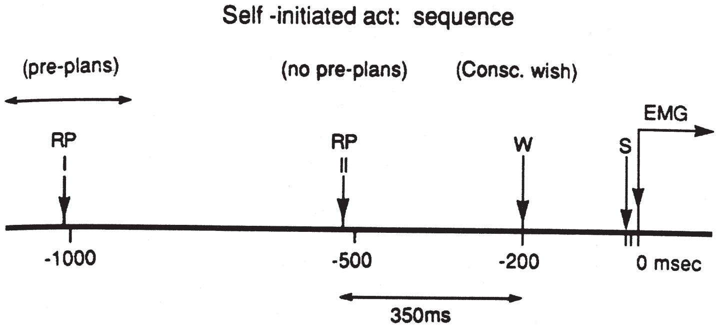


Figure 1.22: The readiness potential (RP) begins first, at about -1050 ms when some pre-planning is reported (RP I), or about -550 ms with spontaneous acts lacking immediate pre-planning (RP II). Subjective awareness of the wish to move (W) appears at about -200 ms, some 350 ms after onset even of RP II. (Figure reproduced from Libet, 1999.)

Libet (1999) concludes that "the volitional process is therefore initiated unconsciously. But the conscious function could still control the outcome; it can veto the act. Free will is therefore not excluded." But this finding is of course contrary to each individual own feeling to consciously initiate voluntary acts.

1.4.2 Conscious will as an illusion

Most of the time in everyday life, we feel we are doing things willfully when we actually do them, and feel we are not doing something when in truth we have not done it. However, some cases remind us that action and the feeling of doing are not inevitably intertwined. The processes of mind that produce the experience of will may be quite distinct from the processes of mind that produce the action itself.

One can think of hypnosis, whose profound effect is the feeling that your actions are happening to you rather than that you are doing them (Lynn, Rhue, and Weekes, 1990). Wegner (2002) also uses the example of table-turning to show that an action can be done without the feeling of having done it. In table turning, a group of people gather around a light table and wait for it to move. Often it would move, sometimes even circling the room or rocking from side to side. Investigations by scientists such as Michael Faraday (using force measurement devices between hands and tables) revealed that the source of the table movement was indeed the participants (Faraday, 1853).

Wegner and Wheatley (1999) were inspired by an ordinary household Ouija board to experimentally test whether people will think they have caused actions when a thought relevant to the action is primed just before the action – whether they actually performed the action or not. People in one experiment were presented with thoughts (e.g. a tape-recorded mention of the word swan) relevant to their action (moving an onscreen cursor to select a picture of a swan).

The movement that participants performed was not in fact their own, as they shared the computer mouse with an experimental confederate who gently forced the action without the participants' knowledge. Nevertheless, when the relevant thought was provided either 1 s or 5 s before the action, participants reported feeling that they acted intentionally in making the movement. On trials when thoughts of the swan were prompted 30 s before the forced action or 1 s afterwards, no inflated experience of will was found (Wegner and Wheatley, 1999).

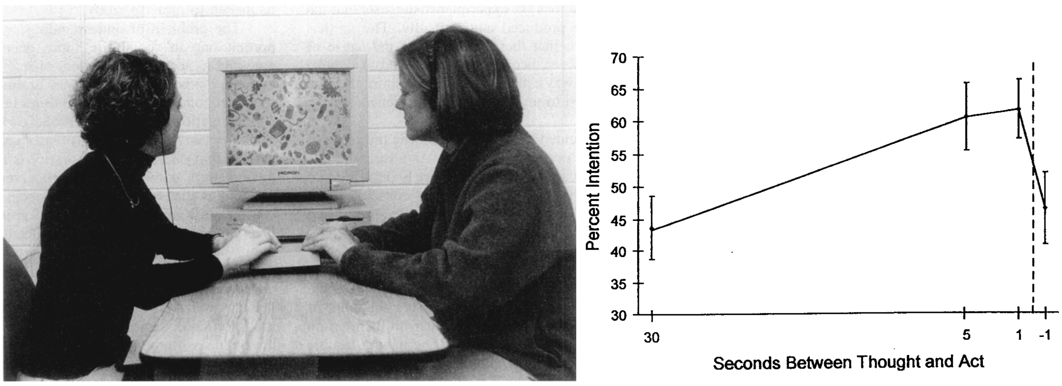


Figure 1.23: On the left, the experimental Ouija board used in the experiment. On the right, the mean percentage of intentionality rated for forced stops on objects primed at different moments before and after the stop. (Figure reproduced from Wegner and Wheatley, 1999.)

This experiment and others have led Wegner (2003) to propose the following theory. The experience of conscious will arise when the person infers an apparent causal path from thought to action (purple arrow). The actual causal paths (green arrows) are not present in the person consciousness. The thought of doing the action, as well as the actual implementation of the action, are caused by unconscious mental events, and these unconscious mental events might be linked to each other directly or through yet other mental or brain processes. Conscious will is experienced as a result of what is apparent, not what is real.

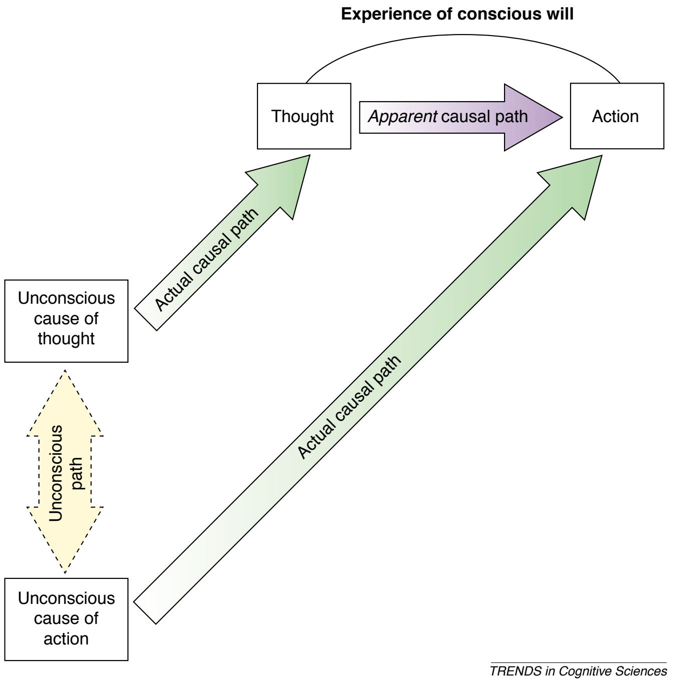

Experience of conscious will
Thought
Apparent causal path
Action
Actual causal path
Unconscious cause of thought
Unconscious path
Actual causal path
Unconscious cause of action
TRENDS in Cognitive Sciences

Figure 1.24: A mechanism to explain illusory agency. (Figure reproduced from Wegner, 2003.)

1.4.3 The importance of believing in free will

Although Wegner (2002) has postulated that our sense of control is a pure illusion, we strongly feel we are in control of our life, and this sense of control is an important part of human psychology (the hypothesis that humans need to feel in control will be developed in the general discussion). What would happen if people came to believe that they cannot exert free will, i.e., that their behavior is the inexorable product of a causal chain set into motion without their own volition?

Believing that outcomes are based on an inborn trait, rather than effort, influences behavior. For instance, Mueller and Dweck (1998) observed 10-year-old children who were told that they had been successful on an initial task either as the result of their intelligence or through their hard work. In a second round, all the children encountered a task that was well beyond their performance level, at which they all failed. When the children were given yet a third task, those who thought their earlier success was due to their intelligence put forth less effort and reported lower enjoyment than those who thought their initial success was due to their own effort. When asked, children praised for intelligence described it as a fixed trait, while children praised for hard work believed it to be subject to improvement.

In two seminal experiments, Vohs and Schooler (2008) have studied whether believing that human behavior is predetermined would encourage cheating. Participants would read either a text that encouraged a belief in determinism (i.e., that portrayed behavior as the consequence of environmental and genetic factors), a neutral text, or a text endorsing free will. They found that weakening free-will beliefs reliably increased cheating: participants who read deterministic statements were less likely to actively prevent the answer to an arithmetic problem from appearing on the computer screen, and overpaid themselves when allowed to take money for each correct answer on a difficult cognitive test.

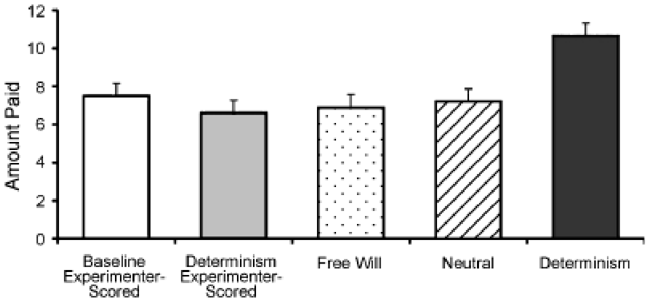


Figure 1.25: Mean amount of money, in dollars, that participants received in five different conditions. Participants in the free-will, neutral, and determinism conditions paid themselves $1 for each answer they claimed to have solved. Participants in the two experimenter-scored conditions were paid according to the true number of solutions. (Figure reproduced from Vohs and Schooler, 2008).

Subsequent work has shown that increasing disbelief in free will contributes to increases in agression and decreases in helpful, prosocial behavior (Baumeister, Masicampo, and DeWall, 2009). A possibility is that the belief that forces outside the self determine behavior drain the motivation to resist the temptation to cheat, inducing a "why bother?" mentality (Vohs and Schooler, 2008). Or perhaps, denying free will simply provides the ultimate excuse to behave as one likes. Sartre (1956) indeed said: "We are always ready to take refuge in a belief in determinism if this freedom weighs upon us or if we need an excuse."

1.4.4 The intentional binding paradigm

As we have seen, explicit judgements of control or agency can be contaminated by a need for excuses, and confounding effects on explicit agency judgements therefore seem inevitable. The intentional binding paradigm offers an implicit measure related to sense of agency, which may be less subject to cognitive biases.

In the first article to report the intentional binding effect, Haggard, Clark, and Kalogeras (2002) used the Libet clock method to study the perceived time of actions and their consequent effects. In baseline conditions, participants either made voluntary actions or listened to the occurrence of an auditory tone (in the absence of action) while they watched a rotating clock hand on a computer screen. They were asked to report the position of the clock hand when they moved or when the tone occurred. In operant conditions, participants made a voluntary key press on every trial, but this time it was followed 250 ms later by an auditory tone.

The authors found that, in operant conditions, the perceived time of their actions was later than in baseline conditions and the perceived time of the tone was earlier than in baseline conditions. Critically, in an identical set of conditions involving involuntary movements induced via transcranial magnetic stimulation (TMS) over the primary motor cortex, the binding effect was reversed such that the interval between action and effect actually increased in 'operant' conditions compared to baseline conditions. The authors speculated that a specific cognitive function of the central nervous system is to bind together critical sensorimotor events that surround voluntary action, and that this function may be crucial for the normal experience of agency.

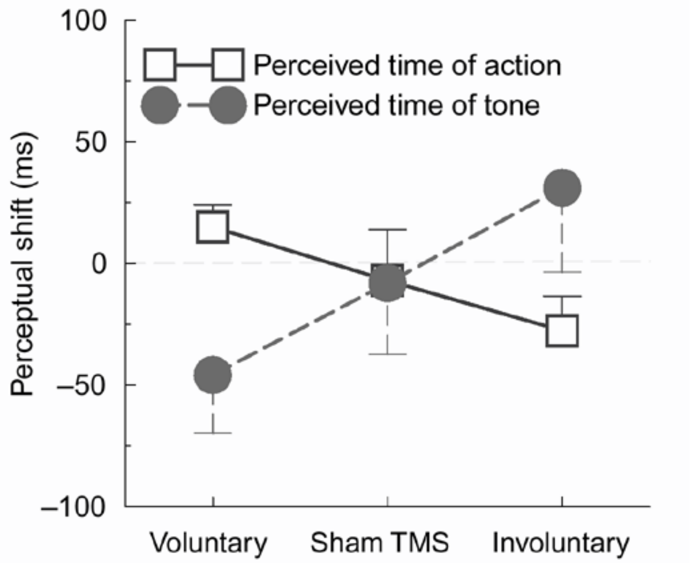


Figure 1.26: Voluntary actions produce binding effects, as awareness of voluntary action shifts later toward a consequent tone, whereas awareness of the tone shifts forward toward the voluntary action that evokes it (left). Neutral events such as sham TMS produce minimal perceptual shifts (middle). Involuntary TMS-induced movements do not sustain binding, but produce repulsion effects in the opposite direction (right). (Figure reproduced from Haggard, Clark, and Kalogeras, 2002).

Since this first report, considerable interest has been generated and a fascinating array of studies has accumulated. More than a decade later, there is compelling evidence supporting a link between intentional binding and sense of agency, although the exact nature of that relationship is yet to be fully understood (Moore and Obhi, 2012).

We have seen that explicit reports of agency or control have shed doubt on the existence of free will, and can be interpreted as a reconstruction of reality. The intentional binding paradigm is thus an important tool to have access to the sense of agency from behavioral measures only.

1.5 The gap between behavior and consciousness

We have seen that there can be a gap between being in control and perceiving to be in control, and that explicit reports can be contaminated by different cognitive processes. Here we will demonstrate that in a variety of learning and decision-making tasks, a general gap has been documented between behavior and explicit reports.

1.5.1 Unconscious conditioning

Conditional responding during simple Pavlovian conditioning is often characterized as a form of implicit memory. The first proof that humans can be unconsciously conditioned came from Morris, Öhman, and Dolan (1998). They measured neural activity in volunteer subjects who were presented with an angry face associated with a burst of white noise. The subjects' awareness of the angry face was sometimes prevented by backward masking with a neutral face. Throughout the experiment, the subjects were required to indicate, by pressing a button, any occurrence of either angry face. Their responses revealed an inability to detect the masked angry faces. Nevertheless, they found a similar significant response in the region of the amygdala to the presentation of the masked and unmasked conditioned faces.

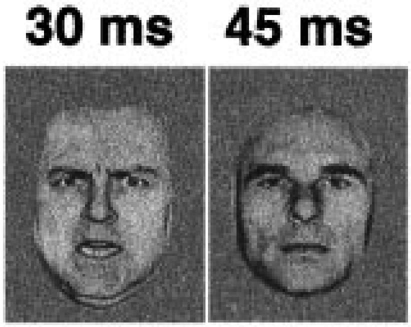


Figure 1.27: The sequence of stimuli used for unconscious Pavlovian conditioning. (Figure reproduced from Morris, Öhman, and Dolan, 1998).

In another study, participants were also exposed to a fear conditioning procedure in which one tone predicted a loud white noise, whereas a second tone was presented alone. The first tone was presented just above or below the perceptual threshold in the different trials. They again found a differential skin conductance response between the two tones, that was present on both perceived and unperceived trials (Knight, Nguyen, and Bandettini, 2003).

Unconscious learning was then demonstrated in instrumental conditioning, showing that unperceived cues could also bias decision making. Pessiglione et al. (2008) used a masking procedure on visual cues, so that participants could not build conscious representations of cue-outcome associations. These unperceived cues were paired with monetary outcomes (+£1, £0 or -£1), depending on whether participants chose the 'Go' or 'NoGo' response. Participants did choose the 'Go' response more frequently following reward predictive cues relative to punishment predictive cues. The uncovered cues were subsequently rated by the participants, and ratings were significantly higher for reward compared to punishment cues, although none of the cues was reported as previously seen.

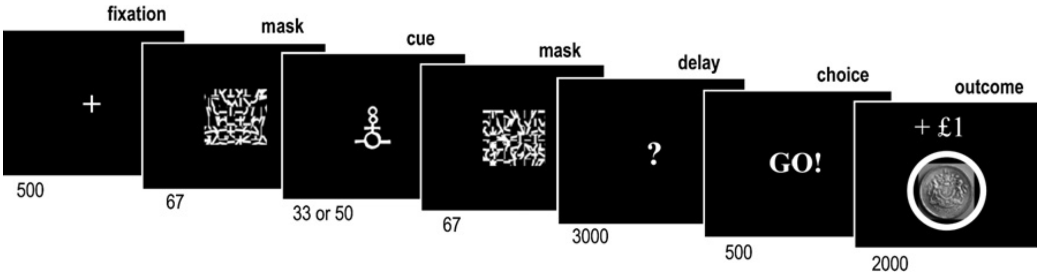


Figure 1.28: The sequence of stimuli used for unconscious instrumental conditioning. (Figure reproduced from Pessiglione et al., 2008).

1.5.2 Implicit sequence learning

Another form of unconscious learning is implicit sequence learning. The seminal study of Nissen and Bullemer (1987) used a simple paradigm: participants responded to a stimulus (an asterisk) occurring at one of four locations with a key located directly below each position. Following a 500 ms response-to-stimulus interval, the next stimulus occurred. The basic design was thus a four-choice, compatible response mapping, serial reaction task. Although not informed of it, some participants were responding to an asterisk moving in a regular, repeating pattern of positions, while others responded to a random order of locations. Subjects in the repeating sequence condition had faster reaction times and made fewer errors than those responding to random information, although they were not consciously aware of a sequence.

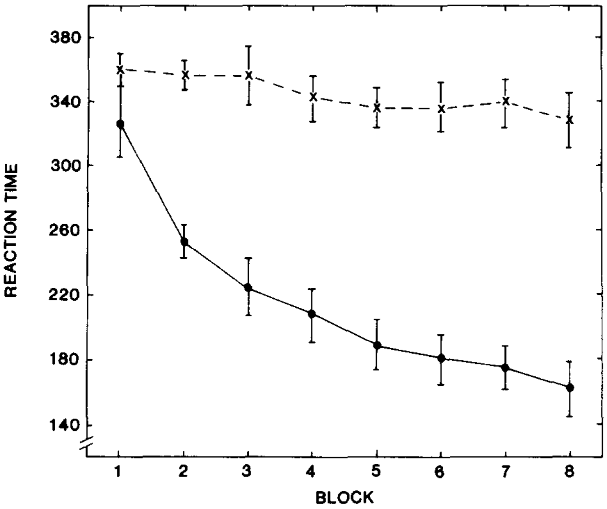


Figure 1.29: The sequence learning phenomenon. (Figure reproduced from Nissen and Bullemer, 1987).

In another study, eight pairs of objects were presented, in which one object in each pair was always correct. Two patients with large medial temporal lobe lesions showed slow and gradual increase of performance in this task, while controls had a perfect score after three days of testing. Crucially these patients were unable to recall or recognize word lists, stories and diagrams (Bayley, Frascino, and Squire, 2005). A study even showed that implicit, automatic learning may be attenuated by explicit memory processes under certain circumstances (Fletcher et al., 2004). Explicit knowledge is thus often not seen as an inherent part of learning sequenced information. Rather, awareness may arise from the interaction of the sequence learning systems with other cognitive systems that then produce conscious knowledge of the sequence (Clegg, DiGirolamo, and Keele, 1998).

Explicit reports can be seen as a pure artificial reconstruction of the real learning process. But this statement actually oversimplifies the interaction between behavior and consciousness. Unconscious instrumental conditioning is known to produce a much smaller performance than conscious instrumental conditioning (Pessiglione et al., 2008). There are also cases in which we actually know more than what our behavior shows. For example, a recent study on mice has shown that task acquisition may be diverging from task expression. The authors analyzed a conditioned response and found a classical and progressive learning curve that is the hallmark of instrumental conditioning. But when testing mice preferences in a different context, they found a all-or-nothing behavior, showing that mice can indeed perform one-shot learning (Kuchibhotla and al., 2018).

A series of experiments have then directly explored the relationship between performance on a cognitive task and the explicit or reportable knowledge associated with that performance. There was no evidence for a positive association between task performance and associated verbalizable knowledge. It seems that subjects are not able to access specific task-related information in a form that will allow them to answer post-task verbal questions. It is possible that whatever is learned during task performance is not verbalizable (Berry and Broadbent, 1984). We can thus envision behavior and conscious reports as two (mostly) independent proxys for cognition.

1.5.3 Feeling free vs. being free to choose

Being free is often described as the possibility of choosing between different things, rather than being forced into an option. But paradoxically when people are faced to a complex choice with multiple options leading to various consequences, they can feel blocked or frustrated rather than free.

In a series of six experiments, Lau, Hiemisch, and Baumeister (2015) studied what factors influenced the feeling of freedom. One experiment compared students having to choose between three, six or nine housing advertisements. The more options the students needed to analyze, the less free they felt. In another experiment, participants had to choose between two job applicants. In one case both applicants were badly fitted for the job, while in another case both were equally competent. Participants felt freer when choosing between two equally good than two equally bad options. Therefore the feeling of choice freedom do not arise from a situation in which a choice can be made between multiple options equally attractive, but rather is due to positive outcomes emerging from or expected from the choice.

Lau, Hiemisch, and Baumeister (2015) concluded that the feeling of freedom essentially differ from what is theoretically seen as freedom. These results can be seen as a consequence of a general well-known phenomenon: choice overload. The choice overload hypothesis states that an increase in the number of options to choose from may lead to adverse consequences such as a decrease in the motivation to choose or the satisfaction with the finally chosen option (see Scheibehenne, Greifeneder, and

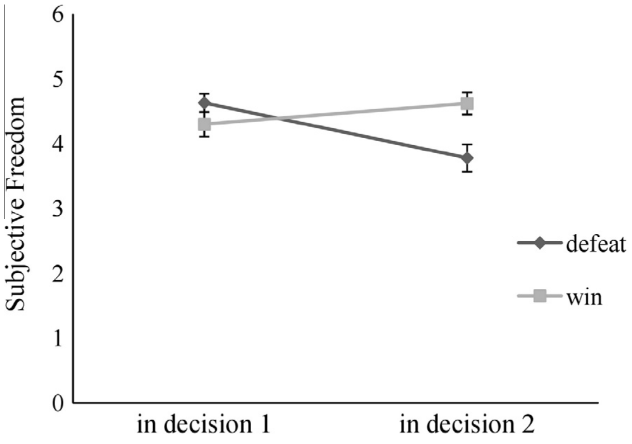


Figure 1.30: Mean experience of freedom for two successive decisions, after a positive or negative outcome occurred after the first choice. The expectancy of obtaining a good result without much effort was a key determinant of the feeling of freedom. (Figure reproduced from Lau, Hiemisch, and Baumeister, 2015).

Todd, 2010 for a review).

In a seminal series of economic experiments, Iyengar and Lepper (2000) revealed the possible negative consequences due to having too much choice. They offered participants a choice between an array of either 6 or 30 chocolates. Participants who chose from the 30 options experienced the choice as more enjoyable but also as more difficult and frustrating. Most intriguingly, though, participants facing the large assortment reported less satisfaction with the chocolates they finally chose than those selecting from the small assortment. This challenges the implicit assumption that the more choice, the better. The reduced feeling of freedom when facing multiple options may thus be an additional consequence of the choice overload phenomenon.

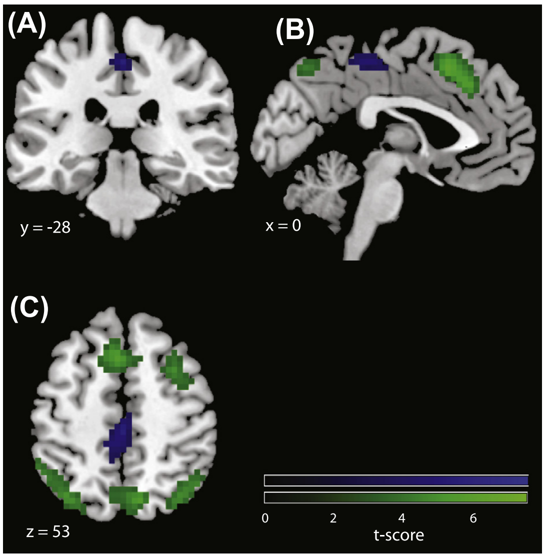


Figure 1.31: The non-overlapping neural correlates of subjective and objective freedom of choice. In blue: the subjective free > subjective instructed contrast; in green: the objective free > objective instructed contrast, both in whole brain. (Figure reproduced from Filevich et al., 2013).

Another study have looked at the difference between being free and feeling free from a neuroimaging perspective. A typical method for studying free choice was used: contrasting free and instructed selection of response alternatives (see Krieghoff et al., 2011 for a review). Filevich et al. (2013) introduced a novel task in which participants had to complete a number sequence with the instruction to make the whole sequence 'look random'. This way depending on the number sequence presented and the personal rule participants followed to make a sequence look random, the experimenters created situations in which the choice of number would feel more free or feel more constrained. BOLD responses for conditions subjectively experienced as free identified a postcentral area, distinct from the areas typically considered to be involved in free action. Their results suggest that the experience of free choice may not derive from brain circuits involved in action selection, but from quite different brain circuits.

Although conscious reports were often used to better characterize control and agency, they could be a purely reconstructive process, dissociated from the cognitive processes underlying behavior.

Chapter 2

Research question

Sense of agency or subjective control depends on the ability to learn and make use of action-outcome contingencies and one of the more classical algorithm to model this learning originates in the field of reinforcement learning. Our aim in this PhD thesis was to study the relationship between control, agency and reinforcement learning processes. As this question is very general, we will focus on three specific problems arising at the interaction of these cognitive processes.

In a first series of experiments, we will study how participants can compute and monitor their control over given outcomes in a changing environment. Our aim was to better understand the computational processes responsible for control perception, and their associated biases. We have seen in the introduction how human participants can be subjected to illusions of control. We thus hypothesized that a model whose updates were built on a by-default control assumption would be more adapted to describe participants' behavior.

In a second series of experiments conducted in collaboration with University College London, we used intentional binding as an implicit proxy to measure participants' feeling of agency in a reversal-learning task. Agency and adaptive response processes have so far been studied almost exclusively separately. Interestingly the emergence of an error in a learning process was shown to increase one's vigilance and cognitive control. We investigated how the implicit feeling of agency is modulated by the error-triggered engagement of cognitive control.

Finally, we investigated whether agency can be the source of some standard sequential decision-making biases, like the choice confirmation bias. In a last set of experiments, the participant could be either an agent or an observer, in a simple instrumental conditioning task. In the agent condition, the subject freely chose between two symbols, whereas in the observer condition, the computer preselected one symbol and the subject was forced to match this choice. Previous experiments have shown that in free-choice, individuals display a choice-confirmation bias. We predicted that a lack of agency at the moment of choice would make the bias disappear.

Subjective reports of control often appeared to be a reconstructive narrative, completely detached from the reality of the task. This is why we will use the tool of cognitive modeling, rather than verbal reports, in this PhD thesis, to uncover the cognitive processes underlying participants' behavior.

Chapter 3

General method

Cognitive modeling is the central, but not necessarily well known, tool used in this PhD thesis. We will first describe here this approach. Theoretical analysis and computational modeling are important ways of characterizing what nervous systems do, determining how they function, and understanding why they operate in particular ways. Cognitive modeling is based on the belief that methods of mathematics, and computer science can provide important insight into cognitive science and psychology (Dayan and Abbott, 2001).

3.1 Cognitive modeling

These past decades, much efforts have been devoted to a model-based approach in neuroscience and psychology, and cognitive modeling has grown considerably in cognitive sciences. The importance of computational models in cognitive sciences and neurosciences is not surprising; because the core function of the brain is to process information in order to guide adaptive behavior, it is particularly useful to formulate cognitive theories in computational terms.

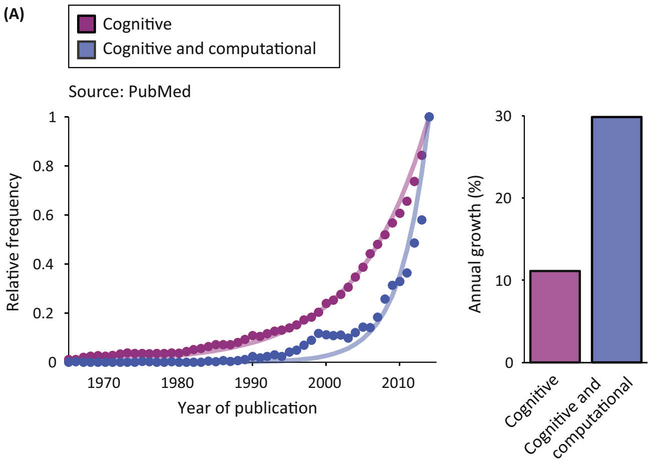


Figure 3.1: The curves on the left show the relative frequency of PubMed entries for ’cognitive’ (in red) and ’cognitive and computational’ (in blue) as a function of the year. Their frequencies are calculated relative to the number of entries of 2014, which are therefore normalized to 1 for both curves. The bars on the left represent the estimated annual growth of the best-fitting exponential curve. (Figure reproduced from Palminteri, Wyart, and Koechlin, 2017.)

3.1.1 Definition

Donoho et al. (2009) have said that originally, there were two scientific methodological branches: *deductive* (for example, mathematics) and *empirical* (for example, statistical data analysis of controlled experiments), but that now, many scientists accept *computation* (for example, large-scale simulation) as the third branch. Weisberg (2007) has rather splited empirical sciences into two main approaches: model-free approaches directly investigate the natural phenomenon of interest, whereas model-based approaches investigate abstract (mathematical) representations of the natural system that are responsible for the empirical phenomenon of interest.

For example, an active area of cognitive modeling is concerned with the question of how we learn to categorize perceptual objects (Medin and Schaffer, 1978). One categorization model is called the prototype model. It postulates that the learner estimates the central tendency from all the examples experienced from within each category during training. When a new target stimulus is presented, the similarity of this target to each category prototype is evaluated, and the category with the most similar prototype is chosen. But according to the exemplar model, the learner memorizes all the examples that are experienced, and the similarity of a new target stimulus is computed to each stored example for each category. These two models differ in terms of the assumptions they make, but they both try to account for a common set of empirical laws to explain categorization.

Cognitive science is concerned with understanding the processes that the brain, especially the human brain, uses to accomplish complex tasks, including perceiving, learning, remembering, thinking, predicting, inference, problem solving, decision making, planning, and moving around the environment. The goal of a cognitive model is to scientifically explain one or more of these basic cognitive processes, or explain how these processes interact (Busemeyer and Diederich, 2010).

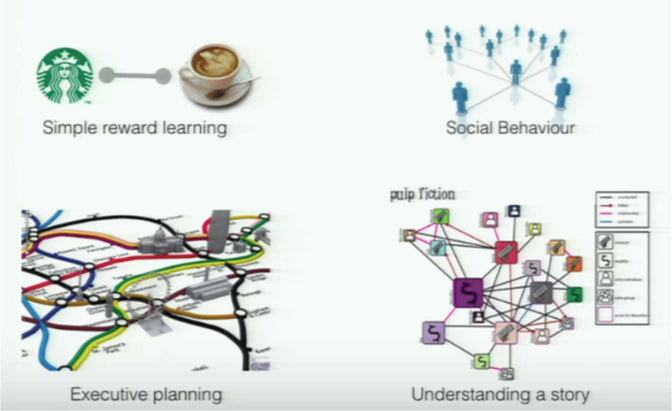


Figure 3.2: Cognitive models can be used in a variety of cognitive science fields, from conditioning to representing social relationships or cognitive maps, or to link together the different characters in a story. (Adapted from Timothy Berhens slides at the Cosyne 2018 symposium.)

But what makes these models cognitive models as opposed to some other kind of models, such as conceptual models, statistical models, or neural models? One hallmark of cognitive models is that they are described in formal mathematical or computer languages. Another hallmark is that they are derived from basic principles of cognition.

3.1.2 Procedure

Busemeyer and Diederich (2010) have described cognitive modeling as a five-stage process:

- A conceptual cognitive theory is reformulated into a mathematical or computer language description.
- As the theory is often insufficient to completely specify a full model, additional ad hoc assumptions need to be made.
- Models almost always contain initially unknown parameters, that need to be estimated from some of the observed data.
- The models are compared with respect to their ability to explain the empirical results.
- Finally one usually needs to start all over again, as model development and testing is actually a never-ending process.

Models always need to be modified or extended to account for newly discovered experimental findings, or in some cases old models need to be discarded for the field to start over. Thus, the modeling process produces an evolution of models that improve and become more powerful over time as the science in a field progresses.

It should be noted that complicated steps in cognitive modeling consist in the addition of ad hoc assumptions and free parameters needed to create a model. Of course theorists always try to minimize their number, to keep the model as simple as possible. One universally recognized heuristic for theory selection is Occam's law of parsimony: "plurality is never to be posited without necessity". This principle dictates that among equally good explanations of data, the less complex explanation should be held as true (Palminteri, Wyart, and Koechlin, 2017).

3.1.3 Perspectives

An advantage of cognitive models over conceptual frameworks is that, by using mathematical or computer languages, cognitive models should guarantee to produce logically valid predictions (provided no calculation or code errors were made). This is not true of conclusions based on intuitively based verbal reasoning that can lead to incorrect conclusions (Busemeyer and Diederich, 2010).

A second important reason for using mathematical or computer models is that they are capable of making precise quantitative predictions. Most researchers would reject a model whose predictions are an order of magnitude off the mark, even though the model makes the correct qualitative or ordinal prediction. One could argue that generic statistical models or empirical curve-fitting models also use formal language and are also capable of generating quantitative predictions. The important difference is that a cognitive model is generalizable: it can be used to derive new predictions for new relationships that go far beyond the original data.

It should be kept in mind that it is meaningless to ask if a model can fit the data or not (Roberts and Pashler, 2000). In fact, all models are deliberately constructed to be simple representations that only capture the essentials of the cognitive systems. The statistician George Box famously said: "All models are wrong but some are useful." (Box, 1979). Indeed a sufficient amount of data will always prove that a model is not true, and no model is able to explain the whole variability of any data set. Therefore cognitive modeling must rely on a comparison between various models.

Thus models are selected on their ability to predict the observed data as a function of their complexity. But the ability of a candidate model to generate a behavioral effect of interest is rarely assessed, although it can be an absolute falsification criterion. Palminteri, Wyart, and Koechlin (2017) have argued that the simulation of candidate models is necessary to falsify models and therefore support the specific claims about cognitive function made by the vast majority of model-based studies.

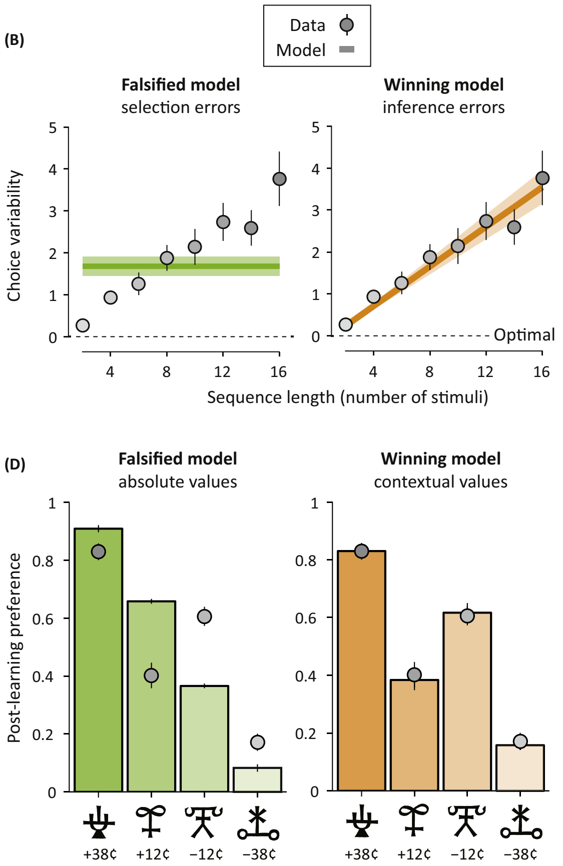


Figure 3.3: Concrete examples of model falsification. *Top panels:* observed (grey dots) and model simulated (colored lines) choice variability in a probabilistic inference task as a function of the sequence length. *Bottom panels:* observed (grey dots) and model simulated (colored bars) post-learning preference as a function of the stimulus value. (Figure reproduced from Palminteri, Wyart, and Koechlin, 2017.)

Now that we have introduced the general framework we will be using in this PhD thesis, we will present one of the most known and used computational framework: the Bayesian or probabilistic framework. This introduction will be short, as we have only used Bayesian models in Study I for comparison purposes.

3.2 Bayesian models

3.2.1 Bayesian probabilities

Bayes theorem is a method of statistical inference that provides a normative way to update a prior belief with incoming evidence. It was named after Reverend Thomas Bayes (1702-1761) who was the first to provide an equation that allows new evidence to update beliefs (Bayes and Price, 1763), then further developed by Pierre-Simon Laplace who published the modern formulation in 1820 (Laplace, 1820).

A probability can give us a measure of how much one believes in something. For example, when one is absolutely sure of something (for example that the sun rises every day), the probability is 1. The use of probability to represent uncertainty is not an ad hoc choice: Cox (1946) showed that if numerical values are used to represent degrees of belief, then a simple set of axioms encoding common sense properties of such beliefs leads uniquely to a set of rules equivalent to the sum and product rules of probability.

Given some phenomenon A and an observation X relative to A, Bayes theorem indicates precisely how much we should update our belief of A given the new observation X:

$$p(A|X) = \frac{p(X|A)p(A)}{p(X)} \tag{3.1}$$

where $p(A)$ is the a priori belief on A, before observing X, $p(X|A)$ is the likelihood of observing X if A is true, and $p(A|X)$ is the a posteriori belief on A taking into account the new observation.

The ideal Bayesian observer is an agent that will use new observations in a normative way. For example, consider the problem of breath cancer. Imagine that 1% of women who participate in routine screening have breast cancer, and that 80% of women with breast cancer will get positive mammographies, while 9.6% of women without breast cancer will also get positive mammographies. Then, according to Bayes rule, a woman who had a positive mammography in a routine screening actually has a probability of only 7.8% of having cancer (Yudkowsky, 2003).

Probabilistic approaches have been increasingly ubiquitous, and widely used, in cognition. Helmholtz (1856) was among the first to propose that the perceptual system executes an "unconscious inference" from sensory stimulations to hypothesize about the environment, but strong experimental evidence in support of this notion has emerged only recently. From knowledge-bases, to perception, to language and motor control, there has been widespread application of sophisticated probabilistic methods in computational modeling (Chater and Oaksford, 2008). These experiments have shown that human behavior is highly consistent with probabilistic reasoning in the sensory (Knill and Richards, 1996) and the motor (Körding and Wolpert, 2004) domain.

For example to quantitatively investigate cue combination, Ernst and Banks (2002) studied how human subjects estimated the width of an object by looking at it and touching it. One could imagine people used the average of the visual and tactile estimates. But Bayes rule would predict that each cue should contribute to the final estimate in proportion to its reliability (or inverse variance). This model behaved very similarly to humans in the task: visual dominance occurred when the variance associated with visual estimation is lower than that associated with tactile estimation.

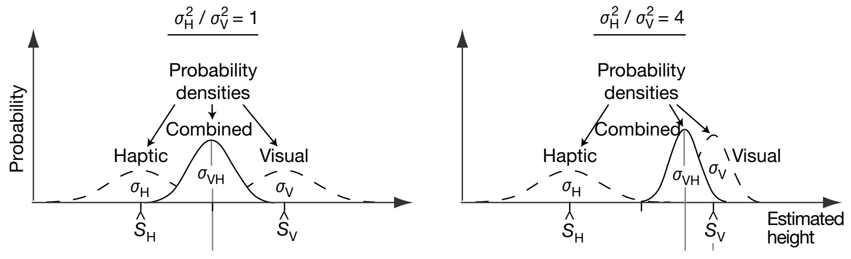


Figure 3.4: Bayesian integration of tactile and visual informations. (Figure reproduced from Ernst and Banks, 2002.)

3.2.2 Bayesian inference

A fundamental notion in Bayesian modeling is the distinction between observed and latent variables. In a Bayesian model, the latent variable distributions are updated based on the values of the observed variables. For example, let us consider a single Gaussian random variable x, whose variance σ^2 is known, and we have to infer the mean μ given a set of N observations. The posterior distribution is given by:

$$p(\mu|x_1, ..., x_N) \propto p(x_1, ..., x_N|\mu)p(\mu) \tag{3.2}$$

If we chose the prior $p(\mu)$ to be Gaussian, it will be conjugate, as the posterior distribution will also be Gaussian. Conjugate priors are often used, as they greatly simplify Bayesian analysis (Gelman et al., 2013).

The mean of the distribution over μ is a parameter controlling a prior, and so it can be viewed as a hyperparameter. Because the value of this hyperparameter may itself be unknown, we can again treat it from a Bayesian perspective by introducing a prior α over the hyperparameter, sometimes called a hyperprior, which is again given by a Gaussian distribution. This construction can be extended in principle to any level, and is an illustration of a hierarchical Bayesian model (Bishop, 2006).

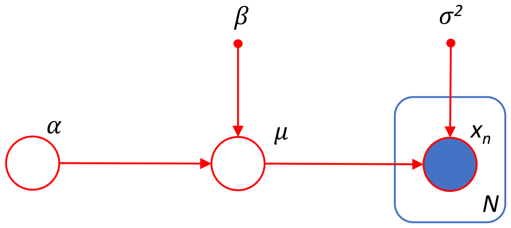


Figure 3.5: The graphical model of a Bayesian hierarchical model. Random variables are represented by empty nodes, and deterministic parameters by smaller solid nodes. The arrows express probabilistic relationship between the nodes. The box labelled N is a plate, representing N nodes of which only a single example x_n is shown explicitly.

Bayesian models have been used in human decision-making, to compute a people's beliefs about the hidden state of the world from observed actions and rewards. In fact, the TD(0) learning rule we presented in the introduction can be seen as a simplified case of the Kalman filter, a Bayesian model that also uses a Temporal Difference learning rule but has additional machinery that determines the learning rate parameter α on a trial-by-trial basis (Behrens et al., 2007).

One key issue in psychology is how to change one's beliefs about the world, and more specifically the amount of influence that unexpected outcomes should have on existing beliefs. The strength of a Bayesian model is that it can distinguish outcomes that are unexpected because of a fundamental change in the environment, from outcomes that are unexpected because of persistent environmental stochasticity. Nassar et al. (2010) have shown that human participants can recognize change points from unexpectedly large prediction errors. This suggests that the brain uses straightforward updating rules that take into account both recent outcomes and prior expectations about higher-order environmental structure.

As we said, Bayesian models allow for a hierarchy in the inferences performed, and this has been used to model cognitive processes. To ensure optimal decision-making, the hidden states and their changes should be probabilistically inferred, but also the rate at which the hidden states will change. Using a hierarchical Bayesian model, Behrens et al. (2007) have showed that human subjects not only adapt their responses when changes occur, but also assess volatility in an optimal manner.

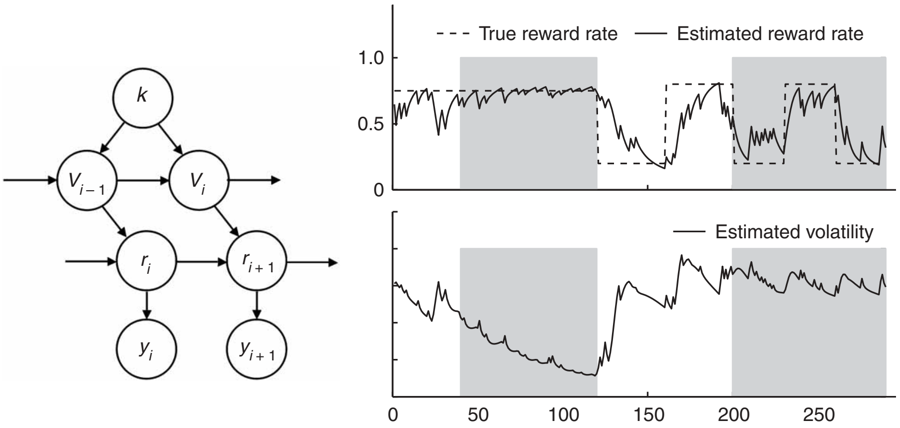


Figure 3.6: *Left panel:* graphical description of the probability-tracking problem. At each trial i, data y_i is observed, which is governed by probability r_i. This probability can change between trials, governed by the volatility, v_i, which can itself change and is governed by control parameter k. The goal of the Bayesian learner is to track these parameters through the course of the experiment, given only the observed data. *Right panel:* the dashed lines show the implemented reward probabilities and volatilities, and the solid lines their Bayesian estimates. (Figure reproduced from Behrens et al., 2007.)

More generally, it is known that cognitive processes can be hierarchically organized in the brain. For example, cognitive control was shown to involve at least three nested levels of processing, implemented in distinct frontal regions (Koechlin, Ody, and Kouneiher, 2003). Another study investigating the architecture of reasoning processes in the prefrontal cortex, have shown that different regions are involved in making probabilistic inferences about the ongoing and the alternative behavioral strategies (Donoso, Collins, and Koechlin, 2014).

3.2.3 Perspectives

Most studies in neuroscience have focused on problems with a small number of variables, all following simple distributions, for which an optimal solution can be easily derived. Examples include integration of two conditionally independent cues, visual search with simple, independent stimuli, and temporal integration of sensory evidence for binary decision-making in a stationary environment. For these tasks, humans and animals often exhibit near-optimal behavior, in the sense that they take into account the uncertainty associated with all signals and combine these signals according to their reliability (Pouget et al., 2013).

Real-life problems, however, are almost always far too complicated to allow for optimal behavior. Optimal behavior requires both full knowledge of the generative model and the ability to perform exact inference, neither of which are possible for most problems of interest.

Given the difficulty of real-world problems, one might imagine that, when confronted with them, the brain no longer relies on a probabilistic approach, but uses instead a set of heuristics (Gigerenzer, Todd, and ABC Research Group, 1999). There are a variety of approximate approaches to hard inference problems. However, whether organisms continue to be probabilistic on hard problems or, alternatively, whether organisms abandon the probabilistic approach altogether when the problems become especially difficult can only be answered experimentally.

The recent advances of cognitive modeling allows us to now predict human learning and decision-making in fine-grained details. Our goal in this PhD thesis is to gain insight on the interaction of control, agency and reinforcement learning, by using model comparison and parameter optimization.

Chapter 4

Study I

4.1 Introduction

We live in a constantly changing world. To adopt a flexible behavior, adapted to new situations, we need to monitor actions bearing consequences in the outside world, and select the most appropriate one. This requires being able to attribute a causal relationship between our actions and external events.

This first series of 3 experiments were built on a modified reversal-learning procedure, in which there was some uncertainty about the identity of the causal agent. To maximize their performance subjects had to continuously monitor their causal influence over the task environment, by discriminating changes that were caused by their own actions from changes that were not.

Our aim was to better understand the computational processes responsible for control perception, and their associated biases.

4.2 Our draft in preparation for *Psychological Review*

Believing in one's power: a counterfactual heuristic for goal-directed control

Valérian Chambon[1,2]*, Héloïse Théro[2]*, Charles Findling[2,3] and Etienne Koechlin[2]

* co-first authors

[1] *Institut Jean Nicod, ENS-EHESS-CNRS, Département d'Etudes Cognitives, PSL University, Paris, France*

[2] *Laboratoire de Neurosciences Cognitives Computationnelles, INSERM-ENS, Département d'Etudes Cognitives, PSL University, Paris, France*

[3] *Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE ParisTech), CREST, 91120 Palaiseau, France*

Corresponding authors: Valérian Chambon (valerian.chambon@ens.fr) and Etienne Koechlin (etienne.koechlin@upmc.fr)
Laboratoire de Neurosciences Cognitives Computationnelles, École normale supérieure, 29 rue d'Ulm, 75005 Paris, France.

July 1, 2018

Most people envision themselves as operant agents, endowed with the capacity to bring about changes in the outside world. This ability to monitor one's own causal power has long been suggested to rest upon a specific model of causal inference, i.e., a model of how our actions causally relate to their consequences. What this model is, and how it may explain departures from optimal inference, e.g., illusory control and self-attribution biases, is still conjecture. To address this question, we designed a series of novel experiments requiring participants to continuously monitor their causal influence over the task environment, through discriminating changes that were caused by their own actions from changes that were not. Comparing different models of choice, we found that participants' behaviour was best explained by a model deriving the consequences of the forgone action from the current action taken, and assuming relative divergence between both. Importantly, this model agrees with the intuitive way of construing causal power as "difference-making": causally efficacious actions are actions that make a difference to the world. We suggest that this model outperformed all competitors because it closely mirrors people's belief in their causal power, a belief that is well suited to learning action-outcome associations in controllable environments. We speculate that this belief may be part of the reason why reflecting upon one's own causal power fundamentally differs from reasoning about external causes.

Keywords

instrumental control; reinforcement learning; Bayesian inference; counterfactual emulation; reference-point dependence

Introduction

Inferring causality, i.e., relating changes in one variable to the causal power of another, is a general, robust, and seemingly built-in, ability of the mammalian brain (Premack, 2007). The ability to draw causal inferences is critical for a wide range of behaviours and functions, from learning and planning to flexibly adapting actions and attitudes to external contingencies (Gopnik, Schulz, and Schulz, 2007). Importantly, such ability may come in two different types, with distinct behavioural advantages, depending on the locus of the cause itself. Thus, if the ability to draw relations between *external* variables is paramount for adaptation and survival, it is even more so when it comes to reckon *oneself* (e.g., one's own choice or action) as the cause of a change in the world.

The ability to envision oneself as an operant agent, endowed with the capacity to bring about changes in the external environment, is classically referred to as "sense of agency" (Haggard and Chambon, 2012). Sense of agency builds on the biologically motivated belief that our actions are causal in nature: they have the power to make things happen, and hence can be implemented as efficient means for pursuing desirable outcomes. A wealth of literature in social and cognitive psychology points toward this representation of one's own causal power as something that is part of our natural endowment (Leotti, Iyengar, and Ochsner, 2010), develops early (Helwig, 2006) and is somewhat irrepressible (Ryan and Deci, 2006). These observations are corroborated by numerous studies showing that people readily experience control over objectively uncontrollable events (Blanco, Matute, and Vadillo, 2011) or are subjected to illusion of control even when no true control exists (Langer, 1975), and even though assuming control does not afford any behavioural advantage or is in fact detrimental to performance (Chambon and Haggard, 2012). The belief in one's own causal power also comes with some advantages: higher levels of instrumental control are associated with greater general health (Bobak et al., 2000), fewer depressive symptoms (Rubenstein, Alloy, and Abramson, 2016), and higher self-esteem (Heckhausen and Schulz, 1995). Conversely, lowered sense of causation makes individuals more vulnerable to external, and potentially damaging, influence (Burger, 2016), and abnormal sense of agency, such as a loss of control over one's actions and thoughts, is long recognized as a key symptom of mental disorders (Schneider, 1959).

Questions have been raised about the function of this belief in one's causal power. The simple fact of exercising control (i.e., of making things happen intentionally) has been suggested to be inherently rewarding (Karsh and Eitam, 2015; see also Zimbardo and Miller, 1958), as reflected by activity in a corticostriatal brain network that overlaps with the neural circuitry involved in reward and motivation processing (e.g., Tricomi, Delgado, and Fiez, 2004; O'Doherty et al., 2004; Bjork and Hommer, 2007). Incidentally, the belief in one's causal power goes with an inherent *need*, so-called "need for control", whereby opportunities to exercise control are preferred over situations with no control, even when exercising control affords no improvement in outcome reward (e.g., Leotti, Iyengar, and Ochsner, 2010; Sharot, De Martino, and Dolan, 2009; Sharot, Shiner, and Dolan, 2010; Suzuki, 1999; Cockburn, Collins, and Frank, 2014; Bown, Read, and Summers, 2003). Exercising control could serve as one of the primary means by which people foster belief in their causal power. Thus, individuals with little experience acting as an effective agent show impaired ability to detect action-outcome contingencies, and hence little belief in their ability to produce desired outcomes (Leotti, Iyengar, and Ochsner, 2010; Maier and Seligman, 2016; Mineka and Hendersen, 1985)[1].

A belief in one's causal power echoes the well-documented need in humans and animals alike to engage in activities simply to experience "competence", that is, a sense of influencing their environment (White, 1959; see also Karsh and Eitam, 2015). Children spontaneously engage in playful exploratory behaviours where the only drive is to effect "changes" in the environment (e.g., putting a finger in a candle, knocking something off a table). Likewise, rats would readily cross an electrified grid (Nissen, 1930) and monkeys perform costly discrimination problems (Butler, 1953) simply for the privilege of exploring and/or interacting with new territory[2]. A persistent inclination to interact with the environment has been suggested to foster action over inaction, which may prove valuable in situations where acting does not satisfy any short-term need. A bias toward action over inaction would thus promote learning of new contingencies by favouring the acquisition of *incidental* associations between actions and action-contingent events. Once learned, these new associations could be then *intentionally used* for pursuing desirable outcomes, i.e. for achieving goal-directed behaviours (Elsner and Hommel, 2001; see also Berlyne, 1950; Berlyne, 1966)[3].

In addition to acquiring new action-outcome contingencies, a belief in one's own causal efficiency may prompt the agent to probe the latent structure of the environment for causal variables. Making decisions based on knowledge of causal variables, rather than based on local changes in the environment only, allows for better anticipating changes in external contingencies and, ultimately, for driving changes in the environment rather than being merely driven by environmental changes

[1] The need to be and feel in control is so strong that individuals do whatever they can to re-establish control when it disappears or is taken away (Brehm, 1966; Brehm and Brehm, 1981). Reestablishment of lost agency can take different forms, from illusory pattern perception to erroneous identification of causal relationship between random or unrelated stimuli. Thus, people experiencing a loss of control are more likely to see images in noise, to form illusory correlations, to perceive conspiracies or to develop superstitions (Whitson and Galinsky, 2008). Such erroneous causal attributions would help restore feelings of control in helplessness individuals by returning the world to a predictable state where "being in control" is the default (Pittman and Pittman, 1980).

[2] An irrepressible tendency for playful and exploratory behaviours parallels Hendrick's "instinct to master", whose aim is merely "pleasure in exercising a function successfully, regardless of its sensual value". This "primary pleasure" would arise when efficient action enables the individual to control and alter his environment (Hendrick, 1943). Interestingly, such exploratory behaviours have been found to be more frequent in younger animals, in which action-outcomes relationships have been less experienced (Siwak, Tapp, and Milgram, 2001).

[3] In a similar vein, it has been suggested that some of our causal beliefs – e.g., control and self-efficacy beliefs – would have evolved to foster the discovery of unpredicted sensory events for which our actions are responsible, hence to reinforce and prioritize those actions that lead to control over the environment (Karsh and Eitam, 2015; Redgrave and Gurney, 2006).

(Koechlin, 2014). Human cognition would be spontaneously framed in such a mode where "being a causal agent" is the default, and self-efficacy beliefs, cognitive instantiations of this default mode (Haggard and Chambon, 2012).

Collectively, the pervasiveness of this default belief in one's causal power (Haggard and Chambon, 2012), the behavioural advantages it affords (Shapiro, Schwartz, and Astin, 1996), and the various functions it underlies (Leotti, Iyengar, and Ochsner, 2010), provide some clues on how human agents calculate and oversee their causal influence on the external world. A belief in the causal effectiveness of one's action is likely to rest upon a specific model of causal inference, i.e., a model of how actions causally relate to their consequences. The general aim of this paper is to uncover what this model is. Crucially, this model should be able to explain how people learn and update their causal influence on a trial-by-trial basis, and make appropriate decisions – such as adjusting behavioural strategies to contingency changes – based on reliable causal estimates. In addition to accounting for the robustness of our everyday inferences, this model should also be simple enough to account for the ease with which human agents calculate action-outcome contingencies – this model should be algorithmically simple. We speculate that simplicity is required to explain how control beliefs can be sustained as a default backdrop to our normal mental life (Chambon and Haggard, 2013). Finally, this model should be endowed with properties that ultimately account for *spontaneous* illusions of control, i.e., for why people readily credit themselves for unrelated events, or perceive control where there is none and act superstitiously in the belief that they are objectively controlling uncontrollable outcomes.

How people track the causal effectiveness of their actions has been a central aim of many empirical investigations, from animal learning to action cognition and personality psychology, through the prism of distinct but related, and often complementary, notions – e.g., intentional causation (Heider, 1958), perceived behavioural control (Rothbaum, Weisz, and Snyder, 1982), self-efficacy (Bandura, 1989), credit assignment (Sutton and Barto, 1998), controllability (Harris, 1996), instrumental learning (Dickinson, 2001), agency (Haggard and Chambon, 2012). Despite the great variety of disciplines concerned, three dominant approaches to instrumental causation can be distinguished, upon which relationships between action and outcome are either:

- retrospectively inferred (associative approach),
- explicitly calculated (generative approach),
- simply emulated (counterfactual approach)[4].

Importantly, each of these approaches draws upon different strategies, with different costs and benefits. Hence, they can be distinguished on several grounds: their *efficiency*, allowing for slow or quick adaption to contingency changes; their *cost*, which makes them likely or unlikely to be implemented by resource-bounded agents; and their *vulnerability* to illusions of control and self-attribution biases. In the next section, we describe typical instances of these three approaches (associative, generative, and counterfactual models), with their respective strengths and weaknesses. Then we turn to formal instantiations of each of these approaches, which we test and compare across in a series of modified probabilistic reversal-learning tasks. We then motivate the development of a computational model from the reinforcement-learning framework – allowing choices to be made online with minimal computational expense –, which we further extend to handle the emulation of unseen (i.e., counterfactual) action-outcome contingencies.

Associative models: causation is about maximizing the expected value of action

One of the dominant views on causation, the associative approach, traces its roots to David Hume (1748). This approach is motivated by the fact that causation is ultimately unobservable, and yet causal relations must be inferred from sensory inputs in some way (see Cheng, 1997; Walsh and Sloman, 2011; Illari, Russo, and Williamson, 2011). According to Hume, three empirical criteria only must be met for characterizing causation: the cause must precede the outcome, the outcome must regularly follow the cause, and both must be spatially and temporally contiguous. Importantly, Hume's definition of causation does not rely on any reference to the mechanism or process connecting events together. Causal relationships are assumed, rather than directly perceived or known, by noticing constant conjunctions between two events, and by retrospectively presuming that a connection underpins their conjunction (Hume, 1748). Ultimately, the associative approach holds that causality is anything but a belief, rooted in our own biological habits – a pure mental construct rather than an objective property of things.

[4] Here we draw upon a classical distinction between associative and generative approaches to causation, according to which causes are "associated" with effects by retrospection or actively "generate" their effects through an operant mechanism (e.g., Cheng, 1997). Strictly speaking, however, associative models in the form of reinforcement-learning (RL) algorithms do also possess a generative model of the world (i.e., an explanation for how observations are generated), whereas counterfactual emulation is a generative mechanism per se (i.e., a mechanism to decide which among several candidate causes has generated the effect). Here, and in what follows, we take the "generative" term in a broader and more liberal sense: generative models are those models drawing on an explicit representation of the generative source (usually in the form of a probability distribution over action outcomes), which can be used to make predictions about future outcome states. Because representation of the generative source is explicit, it is also either complete or approximating the complete solution (that is, an exhaustive representation of all possible action-outcome contingencies). In this sense, generative models are also often normative, i.e., derived from rational principles, and aim at statistical optimality (Gershman, 2015, "normative statistical perspective").

Hume's associative account of causation has inspired various models of causal learning, from contingency models (e.g., Ward and Jenkins, 1965) to the Rescorla and Wagner (1972)'s discrepancy-based learning rule. A typical formulation of the associative approach can be found in studies of instrumental conditioning, whereby causal action-outcome knowledge is acquired through repeated experience with event contingencies, i.e., with repeated associations between some actions (pushing a lever) and motivationally significant events such as rewards (food delivery) (Dickinson, 2001). In the context of reinforcement learning (RL), the Rescorla-Wagner rule formalizes a simple algorithm to account for the acquisition of associative links between event representations on a trial-by-trial basis (Rescorla and Wagner, 1972; Sutton and Barto, 1998). According to this rule, association between action and consequence is learned through experiencing incremental changes in the strength of their link, and learning continues until there is no longer discrepancy between the predicted and the actual consequence of action (Sutton and Barto, 1998). Note that typical RL models do not make use of predictive knowledge (e.g., "cognitive map") about operant actions and their association with relevant outcomes. RL algorithms operate *retrospectively* on experience with previous rewards, by reinforcing actions that were successful in the past – by increasing the propensity to take actions that were followed by a positive reward prediction error (Figure 1A).

RL algorithms present several advantages that can be leveraged to model how people learn and represent their causal power. First, RL algorithms are computationally simple: they typically require a feed-forward mapping of action to predicted consequences (Daw and Dayan, 2014). Their simplicity makes those algorithms robust and adaptive processes that can learn a variety of complex tasks even in uncertain environments (Koechlin, 2016; Gershman, 2015). This simplicity however comes at the cost of inflexibility. Without an explicit representation of instrumental contingencies (including a representation of how alternative, *unchosen* actions impact the environment), an RL agent can only rely on current experience to adjust its behavioural strategies. As a consequence, RL agents require a large amount of experience to learn reliable predictions (Gershman, Markman, and Otto, 2014), and hence may adapt slowly to environments exhibiting action-outcome relationships that change periodically (Koechlin, 2016).

Generative models: causation is about maximizing the dependency between action and effect

Although causal learning exhibits many of the cardinal features of associative processes, there is evidence that human agents do not assess their causal power by simply experiencing (even repeated) conjunctions between what they want, do, and get, as an action effect. Rather, they actively infer causation based on predictive representations of action-outcome relationships, i.e., on internal models of the world that explicitly relates alternative actions to future environmental states (Doya et al., 2002; Daw, Niv, and Dayan, 2005). Drawing upon these internal models, agents do not only notice that effects "follow" their actions: they explicitly represent the generative source that links the action to the effect.

Various models of decision-making have endorsed this "generative" account of causation, from model-based RL to hidden Markov and Bayesian learning models (e.g., Daw, Niv, and Dayan, 2005; Gläscher et al., 2010; Daw and Dayan, 2014; O'Reilly et al., 2013). In a nutshell, generative models hinge on the assumption that the observed data are the realization of one or many hidden variables (the generative source) whose values can be inferred with some degree of certainty, i.e. probabilistically. Crucially, generative models of instrumental causation assume that people have a more or less comprehensive representation of these variables, which can be learned and built up over a history of discrete observable events, or which can be given prior to observation (e.g., O'Reilly et al., 2013). Importantly, these predictive representations of action-outcome relationships can be used to represent the outcome from even *unchosen* actions, and hence to evaluate the different courses of action with respect to the agent's current needs and motivational states.

When inferring action-outcome relationships, the advantages of generative models are multiple. First, their underlying representations of action-outcome mapping make these models statistically efficient: they allow for a potentially optimal use of information derived from experience. Thus, rather than building on the results of the sole action taken, an agent with an accurate estimate of the outcome distribution can potentially evaluate all alternatives at once (Figure 1B). Second, causal relations are computed directly, based on their predictive representations, rather than inferred based on *past* experience with local changes in the stimulus. The generative approach thus allows for more flexibility in adjusting to abrupt or rapid changes in contingencies, as they readily occur in open-ended environments (Koechlin, 2016).

Shortcomings of the generative approach concern both its computational cost and its biological plausibility. Under Bayesian setting, the generative approach assumes that the agent can have an exhaustive representation of all possible states on which the inference is drawn. However, in real-life situations, representing and updating all possible alternatives at once leads to intractable computational costs. This makes the *complete* generative solution unlikely to be implemented by the brain (Eckstein et al., 2004), hence explaining why people often depart from statistically optimal predictions made by normative models (e.g., Waldmann and Walker, 2005; see also Blanco, Matute, and Vadillo, 2011; Gershman, 2015). Interestingly, departures from normative predictions often arise in the form of illusions of control, where people behave superstitiously in the belief that they are controlling uncontrollable outcomes (Langer, 1975), such as those occurring when contrasting instrumental vs. observational learning (Waldmann and Hagmayer, 2005) and naturalistic vs. analytic contexts (Matute, 1996), or while experiencing imposed vs. chosen gamble outcomes (Kool, Getz, and Botvinick, 2013). Generative models have difficulty accounting for such illusions while at the same time failing to address causal problems that human subjects yet easily solve (Sloman and Lagnado, 2015). Therefore, questions have been raised about whether deviations from normativity only capture *approximations* of the true generative solution – e.g., due to limits on the size of working memory or on the quantity of attentional resources – or whether they ask for a rethink of how individuals construe their causal power, i.e., with relatively high efficiency and sustainable computational costs (e.g., Jones and Love, 2011; Markman and Otto, 2011; Bowers and Davis, 2012).

Counterfactual models: causation is about actions that make a difference (with respect to their contingent states)

Associative algorithms describe agents that can adapt to the causal structure of the world with minimal computational expense, while generative models directly infer causation by relying on explicit representations of action-outcome relationships. Hence, associative and generative models of causation stand as opposite extremes on a continuum between statistically efficiency and computational tractability. Importantly, a number of theoretical and empirical works has suggested that counterfactual reasoning might sit in the middle of this continuum.

In the decision-making domain, *counterfactual reasoning* (CF) draws upon representations of what would have happened had another choice been taken (e.g., Boorman, Behrens, and Rushworth, 2011). If a large psychological literature has chronicled the affective consequences of counterfactuals, especially regret, on choice behaviour (Bell, 1982; Coricelli et al., 2005; Roese, 1997), there is also abundant empirical evidence that people generate counterfactuals, i.e., simulate alternative possible events and their outcomes, when they think about causal relations. Thus, when a change to an event leads to a change in the outcome, people rate it as more causal than when a change to the event would not undo the outcome (e.g., Walsh and Sloman, 2011). Similarly, making a counterfactual alternative available strongly influences causal judgments, so that the greater the number of counterfactual alternatives for an event, the more causal this event is perceived (Spellman and Kincannon, 2001; McCloy and Byrne, 2002; Byrne, 2005). While CF plays a role in causal reasoning, it is, however, not equally applied to all types of situations. People are more prone to counterfactual thinking for causal relations that are of behavioural significance to them, such as voluntary actions (see Roese, 1997, for a review). Thus, individuals are more likely to generate counterfactuals when judging causation in situations involving *actions* than inactions ("agency effect", see Byrne, 2002), and *controllable* events (e.g., voluntary choices) than uncontrollable events (e.g., an asthma attack) (Girotto, Legrenzi, and Rizzo, 1991; N'Gbala and Branscombe, 1995). Conversely, decreasing causal power and personal control diminishes the propensity for counterfactual thinking (Scholl and Sassenberg, 2014). Together, these results suggest that there is a close relationship between counterfactual thinking and people's sense of causation for actions under their direct control.

Importantly, the CF account defines a cause as something that makes a *difference* to another event (i.e., the outcome would have been different had another action been performed), endorsing a very intuitive way of construing causation as *difference-making*. In the counterfactual literature, models of causal reasoning (e.g., Pearl, 2000) share this idea with modern instantiations of the associative approach, such as recent accounts based on experienced action-outcome contingency – where contingency is defined as the difference between conditional probabilities, the so-called "ΔP rule" (e.g., Tanaka, Balleine, and O'Doherty, 2008) – and more recently with model-based learning algorithms drawing upon the notion of instrumental divergence (i.e., "Jensen-Shannon divergence"). Instrumental divergence formalizes the causal power of an action as the difference between probabilities of a given outcome in the presence vs. absence of this action (Mistry and Liljeholm, 2016; Liljeholm et al., 2011; Liljeholm et al., 2013). Interestingly, both counterfactual reasoning and instrumental divergence are endowed with the same prior belief about goal-directed actions. They assume that goal-directed actions are instrumental in nature: choosing action A over action B (or choosing to act vs. not acting) *makes a difference in terms of the outcome*. And the greater the action differs with respect to its contingent states (the factual and counterfactual outcomes), the more flexible control the subject has over the environment (Figure 1C).

Importantly, both CF reasoning and instrumental divergence imply being able to *emulate*[5] the outcome associated with the unchosen course of action. Under both views, causal actions are those maximizing the difference between *factual* and *emulated* outcomes. Counterfactual emulation offers several advantages over both the associative and generative approaches. For example, CF makes it possible to learn information from unchosen alternatives without having to incur the costs that taking the alternative course of action would have entailed (Boorman, Behrens, and Rushworth, 2011; Lohrenz et al., 2007; Buchsbaum et al., 2012). Counterfactual emulation is also far less costly than any statistical inference. Unlike generative models that must *learn* causation through considering and

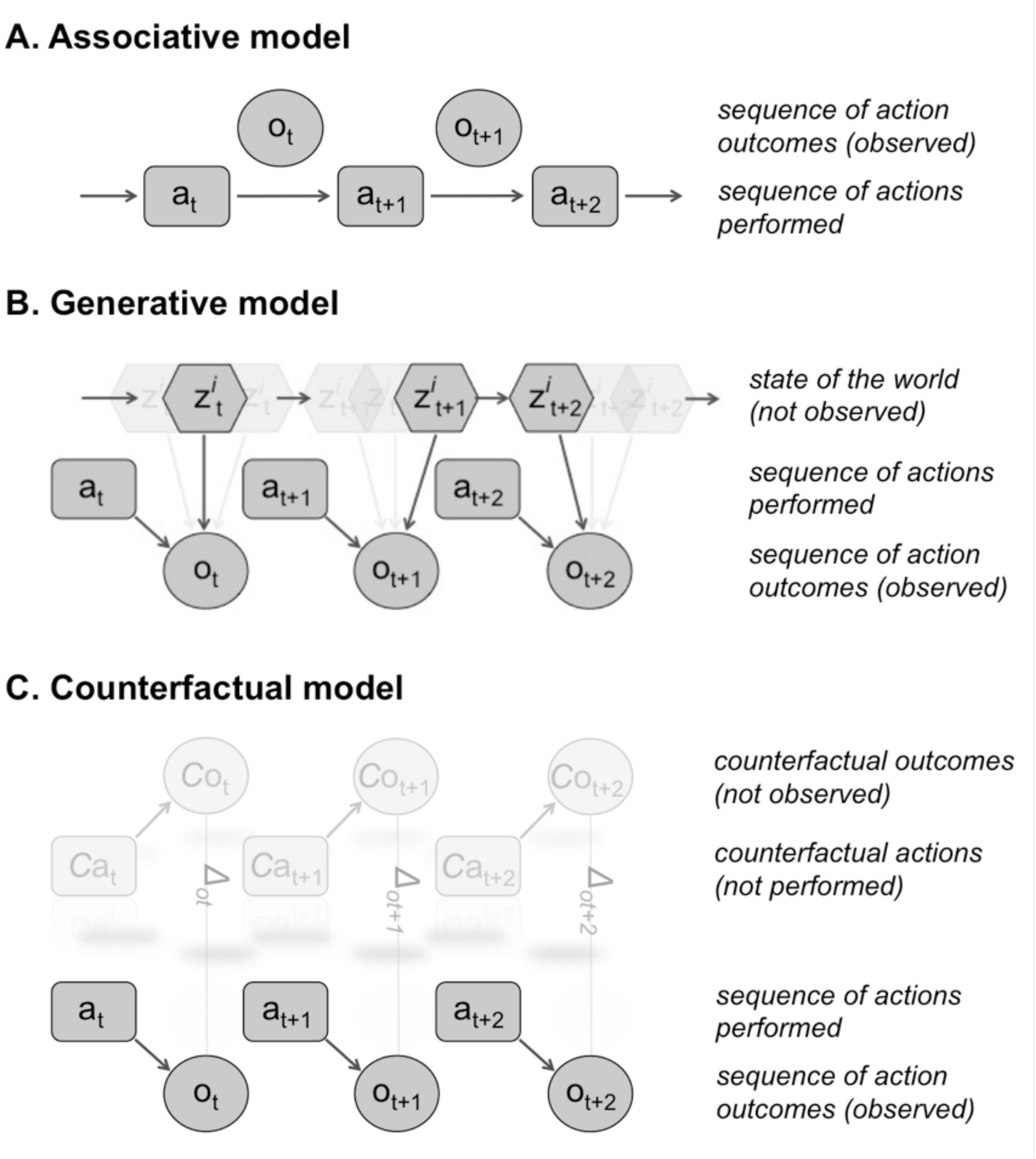


Figure 1: ***Three decision-making and learning models*** *to account for how human subjects infer and monitor their causal power.* ***A. Associative model****: agents associate action (A) and outcome (O) through experiencing repeated event contingencies, and reinforce actions that were successful in the past. Agents learn preferences for actions without ever explicitly learning or reasoning about the (hidden) structure of the environment.* ***B. Generative model****: agents infer action-outcome causal relationships based on an internal "model" of the world (Z; the "generative" source) that explicitly relates actions (A) to future outcomes (O). Generative models can ideally learn all possible hidden states (octagons in transparency) relating the action performed with the observed outcome.* ***C. Counterfactual model****: agents simulate what would have happened (Co) had another action (Ca) been taken. Under the counterfactual view, an action has causal power over an observed outcome if a change in that action (i.e., another action, or no action, is taken) leads a change in the outcome. Ideally, causal actions are those maximising the difference (Δ_o) between factual and counterfactual outcomes.*

updating at once all possible alternative causes, CF assumes causation through a simple prior belief based on difference-making.

[5] Although the term "simulation" is routinely used to describe the process of running (mental) alternatives to the current situation, the specificity of the counterfactual approach is perhaps best captured by a former distinction between *simulation* and *emulation*, as can be found in computer science (e.g., Guruprasad, Ricci, and Lepreau, 2005). A *simulation* represents a target?s behaviour by explicitly modeling its underlying states, usually through a generative model known to best represent the actual states at play. Importantly, however, a simulation does not imply to faithfully mimic the outward behaviour of a target (e.g., a simulation may run faster than real time). *Emulation*, conversely, aims to mimic the observable behaviour, without having to accurately represent its internal states, but with ultimately being able to substitute for the target being emulated (a function ? e.g., face recognition ? emulated by a neural network). Note that emulating an agent, or a function, is useful when one does not exactly know its internal states, or when representing them

Overview of the present study

CF studies have provided convincing evidence that people generate counterfactuals when reasoning about causation (Sloman and Lagnado, 2015, for a review) whereas instrumental divergence provides a learning rule for how people make choice based on maximized divergence (Mistry and Liljeholm, 2016). However, both views have shortcomings. So far CF models of causal reasoning have only been applied to static environments and abstract settings – verbal scenarios or summary descriptions of causal situations –, while studies drawing upon instrumental divergence critically lack of an algorithmic insight into *how* unchosen situations are emulated, and according to which rule (e.g., what value should be assigned to the alternative? how this value can be learned and according to what dynamics? and what should be its update rule?). In this paper, we propose to bridge the gap between these two approaches by building and testing a counterfactual model addressing these issues.

We tested and compared across the performance of this model (hereafter, CF) against various instantiations of the associative and generative classes (hereafter, RL, BM, BC) in a series of tasks where there was some uncertainty about the identity of the causal agent (Figure 2A). The tasks built on a modified reversal-learning procedure (Rolls, 2000) and modelled a dynamic environment where action feedbacks were intrinsically noisy and instrumental or environmental contingencies could change unexpectedly (Figure 2B). To maximize their performance subjects had to continuously monitor their causal influence over the task environment, by discriminating changes that were caused *by their own actions* from changes *that were not*. In turn, discriminating self- from externally-caused outcomes required to track changes in the different statistics manipulated in the task (action-outcome dependency, value, variance) and to flexibly adjust to these changes.

Overview of the experimental paradigm

We ran two distinct experiments, in two different groups of participants. In each experiment, the task consisted in a two slot-machines game (Figure 2A). Thus, on each trial, the participants had to make two distinct choices: (i) first, selecting which of the two machines she wanted to know the result of, then (ii) selecting which of the two buttons (a square or a circle) to press in order to trigger the machines. The order of choice (see Figure 2A: machine then button, or button then machine) was counterbalanced within participants, while the spatial mapping of the task stimuli and the response keys was counterbalanced across participants.

Crucially, the participant was informed that, although she played the two machines simultaneously, she would control *one and only* one machine. Thus, for one machine only, whatever the button pressed (a square or a circle) the average reward was the same, whereas for the other machine, one button (the "best-rewarding" button) gave on average a higher reward than the other (the "least-rewarding" button). Put another way, the chosen button influenced the gains of one machine only (the "controlled" machine), whereas the gains of the other machine (the "non-controlled" machine) were independent of the button pressed by the subject. To maximise her final payoff, the participant had to find out which machine she controlled, that is, the machine for which there was a "best-rewarding" and a "least-rewarding" button.

The participant was informed that she would always win the sum of the gains from *both* machines on each trial. This was to motivate her to track the controlled machine (i.e., the machine for which her choice made a difference) rather than systematically searching for the best-rewarding machine. After a given number of trials, a feedback screen displayed her current payoff, which was graphically represented as the sum of the gains produced by each machine during these last trials.

Gains produced by each bandit machine were drawn from Gaussian probability distributions, of those mean, variance, and "instrumental divergence" varied across conditions. Instrumental divergence refers to the distance between gains distributions associated with each button or machine. This divergence constituted our measure of control. The machine with a *positive* divergence was the *controlled* machine, that is, the machine for which there was a (maximal) difference in the probability distribution of gains associated with each action (e.g., Figure 4A, red and green distributions). Conversely, the *non-controlled* machine was the machine with a *null-divergence*; that is, the machine for which each action was similar with respect to its contingent state (e.g., Figure 4A, grey distribution). Thus, instrumental divergence defines the 'controlled' machine as the machine for which making a choice (e.g., selecting button A vs. B) makes a difference in terms of the outcome, consistently with various accounts of instrumental causation as "difference-making" (e.g., Liljeholm et al., 2013; Walsh and Sloman, 2011; Beebee, Hitchcock, and Price, 2017). It is worth noting that instrumental divergence quantifies the degree to which alternative

accurately would be too demanding.

Simulation and emulation hinge on two different assumptions, which align snugly with the generative and counterfactual approaches to causation, respectively. The generative view assumes that individuals infer causation by simulating the internal process by which hidden states generate observable effects. The process is computationally ruinous, but it may provide an accurate estimate on the likelihood of a candidate cause, given what is observed. The CF view, on the other hand, does not make any reference to the generative source behind observation. Thus, contrary to generative models that simulate *all* possible contingencies from a given situation, CF operates by emulating the unchosen alternative *only*, and by making decision based on variations of some parameters value (e.g., learning rates) when one travels from the real (factual) to the emulated (counterfactual) world (see Lucas and Kemp, 2015).

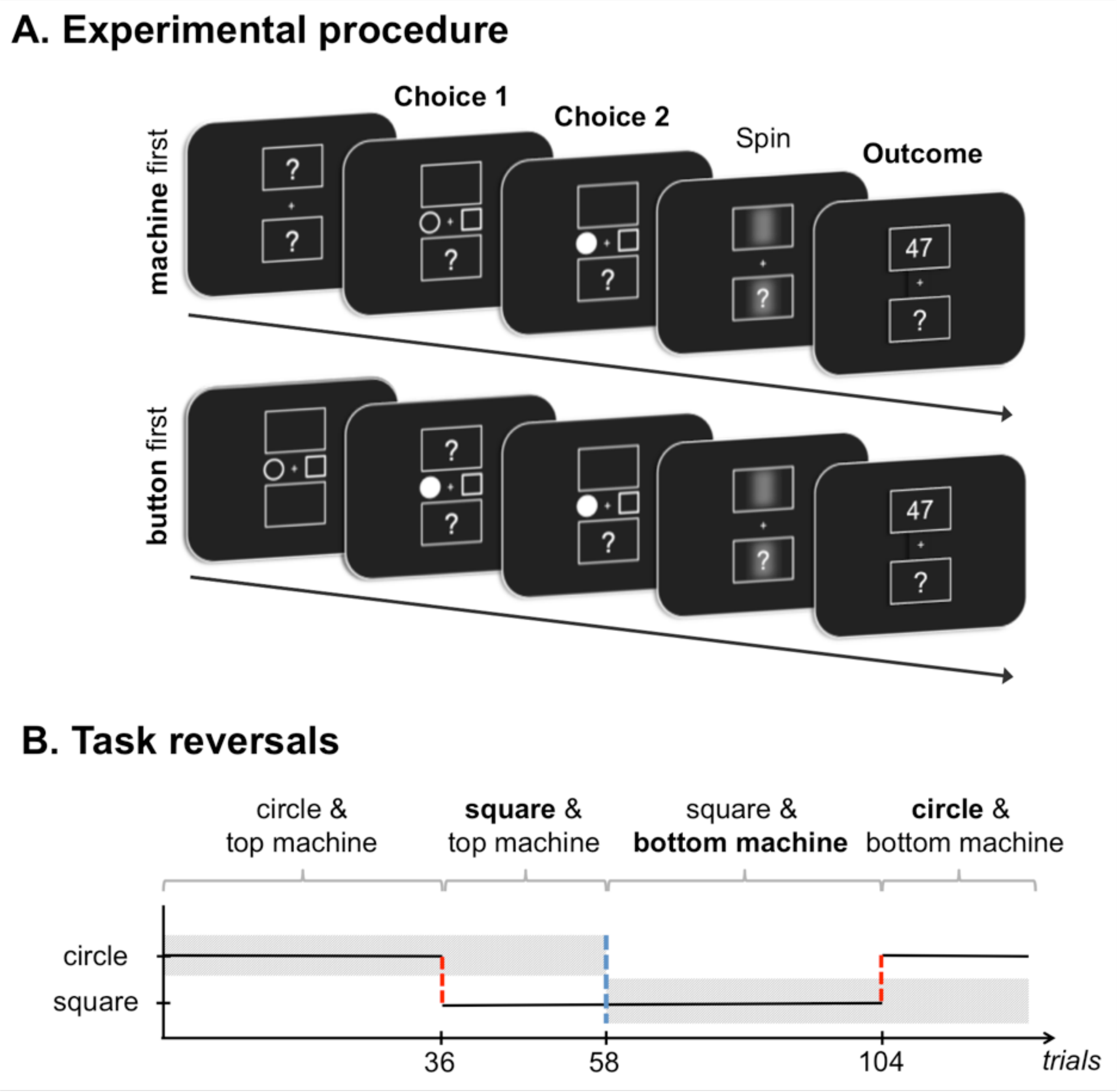


Figure 2: ***Schematic of trial procedure and stimuli.*** ***(A)*** *A trial started with the presentation of two bandit machines above and below a central fixation. In one half of the blocks, the subject had to first select a machine (here, the top machine,* Choice 1*, top panel) and then a button (here, the left button,* Choice 2*, top panel), and conversely in the other half (button, then machine; see bottom panel). Note that only the gains of the selected machine were displayed at the end of the trial. Each trial lasted approximately 3s.* ***(B) Schematic of reversals during the task.*** *The solid line represents the best button during the ongoing block, whereas the grey rectangles represent the location of either (i) the controlled machine (expt. 1: dependency session), (ii) the best-rewarding machine (expt. 1: value session), or (iii) the low-variable machine (expt. 1: variance session). The vertical red dashed lines signal a reversal on the best-rewarding button (circle to square, or the converse) whereas the vertical blue dashed line signals a machine reversal. In all experiments, "button" or "machine" reversals occurred after a variable number of trials.*

actions differ with respect to contingent states, and hence is formally equivalent to another highly related information theoretic measure, mutual information, which quantifies the statistical dependency between an action and a subsequent event (see Liljeholm et al., 2013).

Finally, participants were informed that unpredictable reversals could occur during the task so that either buttons or machines reversed unpredictably from time to time (e.g., the best-rewarding button became the least-rewarding button, or the controlled machine became the non-controlled machine) (Figure 2B). Participants were thus explicitly asked to pay attention to the relationship between their gains and their choices so as to identify these reversals as fast as possible and to adapt their choices accordingly.

Importantly, the experimental conditions differed in how these reversals were implemented. Thus, depending on the condition within each experiment, participants had to monitor reversals in either:

- the statistical dependency between their action and the resulting outcome,
- the rewarding value of the outcomes produced by each machine,
- the variability of these outcomes over time.

The first experiment tested the influence of these statistics on the participant's choice separately, i.e., within independent experimental sessions. In this experiment, either *explicit* (Expt. 1a) or *implicit* (Expt. 1b) instructions were given to participants about their actual control over the task. The second experiment (Expt. 2) implemented the same procedure but controlled for *interaction effects* between the 3 statistics manipulated (dependency, value, variability) by employing a full factorial design in which these statistics were systematically crossed.

Modelling

In both experiments, four classes of models were built and fitted to participant's choices: (i) a simple reinforcement learning model (RL), (ii) a counterfactual learning model (CF) built in a model-free reinforcement learning framework, and two generative models whose aim was to learn the task environment correctly by searching for either (iii) the "best-rewarding" state (Bayesian-maximizer, BM) or (iv) the "controlled" state (Bayesian-controller, BC) in the task environment. Each of these four models draws on different assumptions about how subjects' beliefs are formed and updated on a trial-by-trial basis, and hence makes different predictions on how choices are made based on these beliefs.

In all four models, the *two-stage* decision process was concatenated into one single decision made between the four possible combinations of machines and buttons. We did so because reaction times suggested that the two successive choices (machine then button, or button then machine) were chunked into one unique choice made between four action sequences. Indeed, in all tasks reaction times for choices were significantly slower for the first choice made, whether this choice was a button (paired t-tests, all experiments: all $t_{15/25} < -2.91$, all $p < 0.007$) or a machine (all $t_{15/25} < -11.1$, all $p < 0.001$). In these conditions, it has been shown that modelling two successive choices as one unique decision better predict the participants' data (Dezfouli and Balleine, 2013; Solway and Botvinick, 2015).

In all four models, each of the 4 possible actions made by the participant on each trial (choosing between 2 buttons × 2 machines) was associated with either an *action value* for both RL and CF models (Figure 3), or with *beliefs* (indexing the probability to be in one particular state among all possible states) for the generative models (Figure 4). All four models went through the same two steps on each trial. The first step consisted in updating the internal value or the beliefs associated with each of the 4 possible actions, depending on the outcome obtained in the previous trial. The updating rule was different between models (see below). The resulting internal values or beliefs were then used to compute the probability to choose one action over its 3 alternatives. The second step consisted in making a choice based on either internal values or beliefs, using a non-deterministic (softmax) decision rule (see below, "Action selection").

RL model

Each of the 4 possible actions was associated with an internal value (Sutton and Barto, 1998), also called an action-value (Figure 3, top panel). The values themselves are hidden, but are thought to drive choices between alternatives actions. Specifically, the model draws upon the notion of prediction error (δ), which measures the discrepancy between actual outcome value, called reward (R) here, and the expected outcome for the chosen action (i.e., the chosen value) at time step t:

$$\delta(t) = R(t) - V_{chosen}(t)$$

According to the Rescorla and Wagner (1972)'s rule, such prediction error is used to update the value of the chosen action, as follows:

$$V_{chosen}(t+1) = V_{chosen}(t) + \alpha_F \times \delta(t)$$

α_F is a fitted parameter capturing the rate at which prediction error updates the action values, thus it is called the (factual) learning rate. Action values represent the reward value expected for choosing this particular action. Here the action values associated with the three *unchosen* actions are kept constant (i.e., they are not updated):

$$V_{unchosen}(t+1) = V_{unchosen}(t)$$

CF model

Contrary to typical RL, values of the *unchosen* actions (i.e., counterfactuals) were explicitly updated in the CF model, and this update was performed according to a specific dynamic (e.g., learning rate). Note that counterfactual rewards were not experienced or seen, and therefore must be somehow inferred by the participant. CF models assume that such inference requires to emulate the unchosen action, as *if* it was effectively taken, and to derive the corresponding counterfactual outcome from it (Figure 3, bottom panel). Some uncertainty remains about how to model the emulation process. Converging evidence from reinforcement comparison methods (Sutton, 1984; Dayan, 1991; Kaelbling, Littman, and Moore, 1996) and behavioural economics (Palminteri et al., 2015; Denrell,

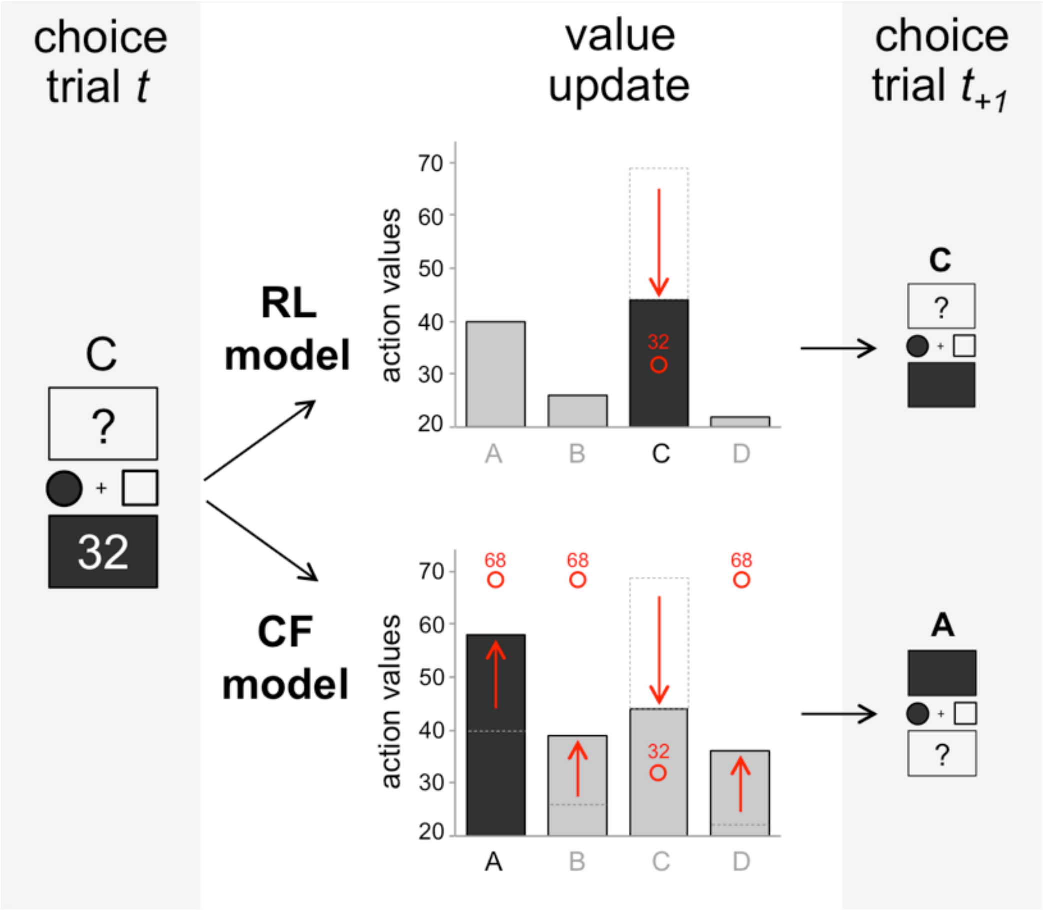


Figure 3: ***Schematic of the two-stages decision process in RL and CF models.*** *On trial t, the circle button and the bottom machine are chosen, and '32' is obtained as a reward. RL (top) and CF (bottom) models differ in how action values are updated. While both models use the current reward to update the value of the chosen action through the Rescorla-Wagner (R-W) rule, only the CF model updates the value of the* unchosen *actions. The CF model derives a fictive counterfactual outcome (here, '68') from the actual outcome ('32'), which it mirrors through a reference point approximating the mean of the underlying reward distribution. The counterfactual outcome is then used to update the value of the unchosen actions through the classical R-W rule. Importantly, these different updates can lead the two models to make different choices: on the next trial, the RL model chooses the circle button and the bottom machine (choice C), while the CF model chooses the circle button and the top machine (choice A). Note that the figure shows a reversal in contingency (C is no longer the best valued action). As can be seen, the CF model adapts quickly to the reversal (it now chooses A), whereas the RL model sticks to the same action (it keeps choosing C as before).*

2015; Burke et al., 2016) suggests that people always make decisions relative to a context-dependent reference. When the context is a (binary or continuous) distribution of gain and losses, this reference approximates the mean of the distribution (Palminteri et al., 2015; Kahneman and Miller, 1986). Interestingly the mean is an important, often optimal, operator that allows for minimizing error in error-prone situations, i.e., under uncertainty (De Gardelle and Summerfield, 2011). In the following, counterfactual rewards were thus inferred based on a simple contextual rule. The counterfactual reward (R_{CF}) was derived from the actual reward, which it mirrored through a reference point (P) approximating the mean of the underlying generative distribution. The value of this reference (or "context value", Palminteri et al., 2015) was separately fitted, rather than fixed or learned from reward history, in each participant:

$$R_{CF}(t) = 2 \times P - R(t)$$

According to this rule, when participants obtained a high reward ("high" being defined as being above the reference), the counterfactual reward associated with the unchosen action was inferred as being a "low" reward (i.e., below the reference reward), and the probability to stay with the same action on next trial increased. Conversely, when the obtained reward was low the counterfactual reward was inferred as being "high", and the probability to switch action on next trial increased. The emulated counterfactual reward thus allowed for computing a *counterfactual prediction error* (δ_{CF}) and a *counterfactual learning rate* (α_{CF}), which was used to update the value of the unchosen actions according to a generalized version of the Rescorla and Wagner's rule, as follows:

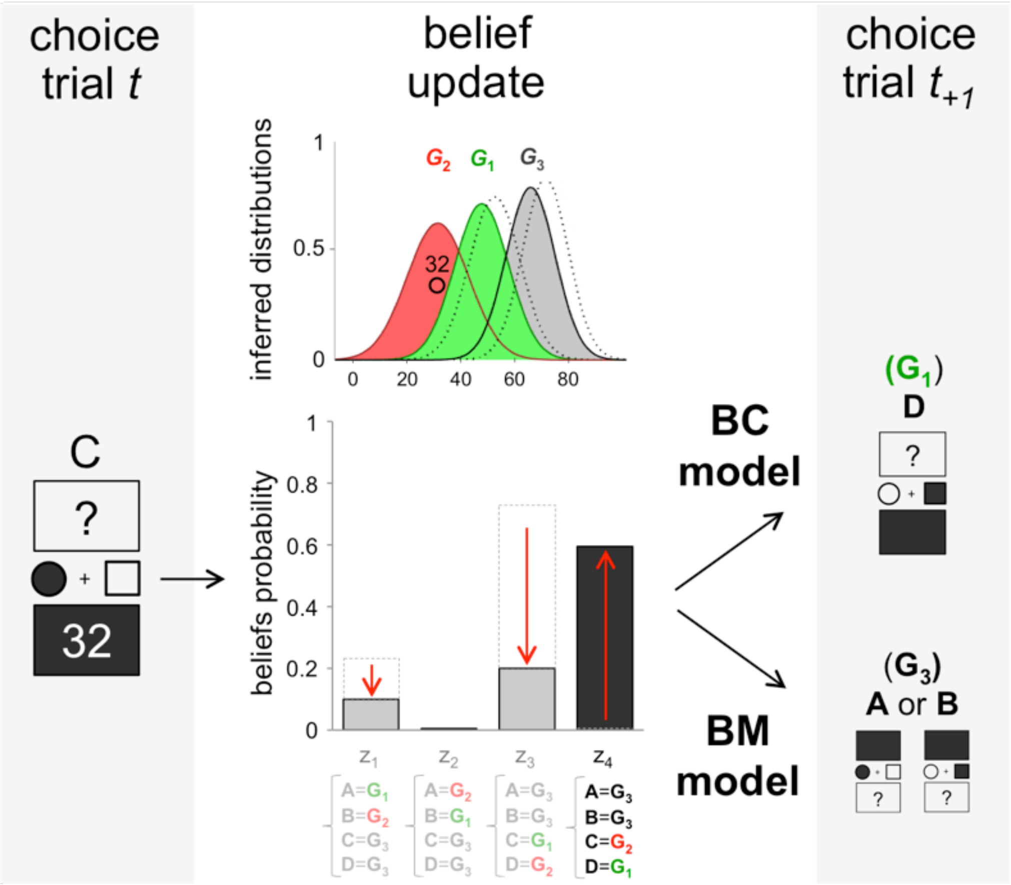


Figure 4: ***Schematic of the two-stages decision process in BC and BM models.*** *On trial t, the circle button and the bottom machine are chosen, and ?32? is obtained as a reward. Both BC and BM models infer the current state of the world (Z_1, Z_2, Z_3, or Z_4, bottom panel) based on the inferred reward distributions (G_1, G_2, G_3, top panel), the volatility parameter and the past history of actions and rewards. The models also update the mean and precision of the three underlying distributions (from dashed to solid distributions, top panel). Because the two models aim to maximize different statistics (control for BC, value for BM), they end up choosing different actions from the state inferred (here, Z4). Thus, on the next trial, the BC model chooses the left button and the bottom machine (D = the best-rewarding action of the controlled machine), while the BM model chooses either A or B, i.e., the current* best-rewarding *machine.*

$$\delta_{CF}(t) = R_{CF}(t) - V_{unchosen}(t)$$
$$V_{unchosen}(t+1) = V_{unchosen}(t) + \alpha_{CF} \times \delta_{CF}(t)$$

Note that because participants chose between four possible actions, there were necessarily *three* unchosen actions for each choice made: (i) the unchosen button associated with the chosen machine, (ii) the chosen button associated with the unchosen machine, and (iii) the unchosen button associated with the unchosen machine. Hence, the model was endowed with 3 counterfactual learning rates (α_{CF1}, α_{CF2} and α_{CF3}), which were fitted in each participant separately.

Generative model

The generative model was a Bayesian learner that updated beliefs, and not values, associated with each possible action, on each trial. Here, a belief referred to a *probability* for an action to be in a given state (Figure 4, and Appendix A, "Generative model"). Instructions that were explicitly given to participants defined four possible states, associated with three generative distributions (G):

- the state associated with having selected the best-rewarding button of the controlled machine (G_1),
- the state associated with having selected the least-rewarding button of the controlled machine (G_2),
- the state associated with having selected the non-controlled machine (G_3).

On each trial, the model aimed to infer the correct state/action pair, i.e., to infer which among the three possible distributions generated the observed outcome, given the button pressed. The model then updated its belief about all state/action pairs, together with the parameters (mean, standard-deviation) of each generative distribution, given the new observations. Our model was implemented with a specific task structure defining the number of possible states (the three generative distributions), actions (the four possible actions), and hidden variables to describe them (e.g., the mean and standard-deviation of the generative distributions). The model assumed the generative distributions to be Gaussian with fixed mean and standard deviation. On each trial, the mean and variance of each generative distribution was inferred by the model, based on the history of observations, through Bayesian inference (see Appendix A, "Generative model", for details). As reversals between actions occurred, the model also needed to infer a volatility parameter, the volatility being the probability for the states to reverse between actions. Thus, on each trial, the Bayesian models needed to infer a set of 7 parameters: the three Gaussian means, the three Gaussian variances, and the volatility parameter (Figure S1).

To test how participants interpreted our instructions, we built two different Bayesian models: a Bayesian-controller (BC), and a Bayesian-maximizer (BM) model. The first model (BC) preferentially selected the action which it believed was associated with the controlled machine (i.e., the model made choices based on a control belief, see Figure 4, top panel) while the second model (BM) preferentially selected the action associated with the best-rewarding Gaussian, irrespective of whether this Gaussian was or was not associated with the controlled machine (i.e., the model made choices based on the magnitude of the reward see Figure 4, bottom panel).

Action selection

Across all four models, action and belief values were used to drive action selection. On each trial, this selection was made through a softmax rule, based on either updated action-values or beliefs (Daw et al., 2006). Under this rule, one action is stochastically selected according to the difference between each action's expected value:

$$P_1 = \frac{e^{\beta \times V_1(t) + \rho_m \times c_{m,1}(t) + \rho_b \times c_{b,1}(t)}}{e^{\sum_i \beta \times V_i(t) + \rho_m \times c_{m,i}(t) + \rho_b \times c_{b,i}(t)}}$$

where i enumerates over all possible choices and $c_{m,i}$ and $c_{b,i}$ were defined as the stickiness to the previous machine or button choice, irrespective of the reward history:

$$c_{m,i}(t) = \begin{cases} \text{1 if the same machine was chosen on t-1} \\ \text{0 otherwise} \end{cases}$$

$$c_{b,i}(t) = \begin{cases} \text{1 if the same button was chosen on t-1} \\ \text{0 otherwise} \end{cases}$$

The exploitation intensity parameter β is fitted and represents the strength of the action values or beliefs on action selection. The parameters ρ_m and ρ_b capture the participant's propensity to perseverate with their action choice, which cannot be explained by reward history (Lau and Glimcher, 2005).

Parameters fitting

Model parameters were fitted based on participants' actions. Model fitting was performed separately for each participant and each condition. The best parameters were those maximising the log-likelihood (LLH), defined as the sum of the log of the model's fit to participants' action choices. Thus, LLH close to 0 indicates a good model fit. To test the different possible combinations of parameters, we used a slice sampling procedure (Bishop, 2006). More specifically, using three different starting points drawn from uniform distributions for each parameter, we performed 100,000 iterations of a gradient ascent algorithm to converge on the set of parameters that best fitted the data.

All four models shared the same three parameters: the perseveration biases ρ_m and ρ_b, and the exploitation intensity parameter β. The two Bayesian models (BC and BM) had no additional parameter to fit, since the parameters used to compute the beliefs were inferred. The RL and CF models shared the learning rate parameter α_F, but the CF model had 4 additional parameters: the three counterfactual learning rates (α_{CF1}, α_{CF2} and α_{CF3}), and the reference point (P). To account for the risk of overfitting, a relative quality-of-fit metric, the Bayesian Information Criterion (BIC), was also computed. The BIC penalizes models with a high number of parameters:

$$BIC = k \times log(N) - 2 \times LLH$$

with k being the number of parameters and N the number of trials.

BIC values were compared between our four models (RL, CF, BM and BC). As an approximation of the model evidence, individual BICs were fed into the MBB-VB toolbox (Daunizeau, Adam, and Rigoux, 2014), a procedure that estimates how likely it is that a specific model generates the data of a randomly chosen subject (the posterior probability of a model, PP), as well as the probability that a given model fits the data better than all other models in the set (exceedance probability, XP).

Choice simulation

The four resulting models (RL, CF, BM and BC) were simulated with the best-fitting parameters, and they underwent the same experimental conditions as participants did. On each trial, the outcome given to the model was the one associated with the model's choice, and not the participant's. Simulations were used to provide aggregated measures of models' performance (e.g., Figure 5B) but also to compare trial-by-trial choice sequence after reversal across models (e.g., Figure 6A).

Experiment 1: method

Participants

16 participants (8 females, age between 20 and 33 years-old) took part in Experiment 1. They provided written informed consent prior to the experiment and were all paid 20 euros for each experimental session completed. No participants had a history of neurological or psychiatric disorder, and all had a normal or corrected-to-normal vision. The experiment was approved by the local ethics review board (CCP C07-28). Participants were informed about the general procedure of the experiment through detailed written instructions.

Stimuli and trial structure

On one half of the experiment, the first choice consisted in selecting a machine, then selecting a button, and the reverse on the other half (button first, then machine). The order of choice was counterbalanced within participants.

When the first choice was about the machine, a typical trial started with the presentation of two machines above and below a central fixation. Each machine was filled with a question mark (see Figure 2A, top panel, "machine first"). Participants had 700ms to make their choice. Once the selection made, the question mark within the chosen machine disappeared. After a 500ms delay two buttons (a square and a circle) appeared on both sides of the central fixation. Again, the participant had 700ms to choose one button by pressing the corresponding key. The chosen button was then filled with white to confirm the participant's key press. Once the choice made, the two slot machines spinned for 200ms. The gain corresponding to the chosen button then appeared in the chosen machine for 800ms. If the participants did not press a key within the 700ms delay, or if the wrong key was pressed, the trial was "missed", and the next trial started. Each trial lasted approximately 3s.

The same timeline applied on trials where the first choice to make was about selecting a button (Figure 2A, bottom panel, "button first"). As mentioned above, the spatial mapping of task stimuli and response keys was counterbalanced across the 16 participants: for half of them the machines were positioned on a vertical line whereas the buttons were on a horizontal line (as represented in Figure 2A). This mapping was reversed for the other half, as were the response keys.

Participants were informed that they would always win the sum of the gains from both machines on each trial. Thus, every 208 trials, a feedback screen displayed the participant's current payoff, which was graphically represented as the sum of the average gains produced by each machine during the last 208 trials. In total, a session consisted in 832 trials. Each session was preceded by a short training (64 trials).

Experimental sessions

Participants completed three sessions, each carried out in a different day and lasting approximately one hour. Each session required participants to track occasional changes in the structure of the task environment, and to adjust their choices according to whether these changes related to either i) the statistical dependency between the option chosen and the subsequent outcome, or ii) the value or iii) the variability of the outcomes produced by each machine.

Thus, each session was defined according to the type of statistic manipulated in the task:

1. The **statistical dependency** between the action made and the resulting outcome was manipulated in the first experimental session. This session implemented a "controlled" (divergent) machine for which each button led to a different outcome, and a "non-controlled" (non-divergent) machine for which the reward was the same, regardless of the button pressed. For the controlled machine the gains associated with the best- and least-rewarding buttons were discretized rewards drawn from Gaussian probability distributions with identical variance (SD) but different means (see Figure 5A, left panel, green and red distributions, means = 58 and 42, SD = 10, respectively). For the non-controlled machine, the gains associated with both buttons were drawn from the same Gaussian (Figure 5A, left panel, grey distribution, mean = 50).
2. The **value** of each machine was manipulated in the second session by implementing a machine that was on average more rewarding than the other, while keeping the two machines non-divergent. Thus, regardless of the button pressed, outcomes from each machine were drawn from Gaussians with identical variance but different means, such that the mean value of one machine (the "best-rewarding machine", see Figure 5A, middle panel, light grey, mean = 58, SD = 10) was systematically higher than the other (the "least-rewarding machine", dark grey, mean = 42, SD = 10).
3. In a third session the **variance** of each machine was manipulated by making the gains from one machine more variable than the other, while keeping the two machines non-divergent. Thus, regardless of the button pressed, the two machines were associated with Gaussian distributions that had the same mean but different variance (low- and high-variable machines: mean = 50, SD = 5 and 15, respectively) (see Figure 5A, right panel).

We were first interested in assessing (i) whether, and how, the three statistics manipulated could influence participants' control beliefs, and second (ii) whether and how well each class of models could account for this influence on participants' choice behaviours. To assess independently the influence of the "value" and

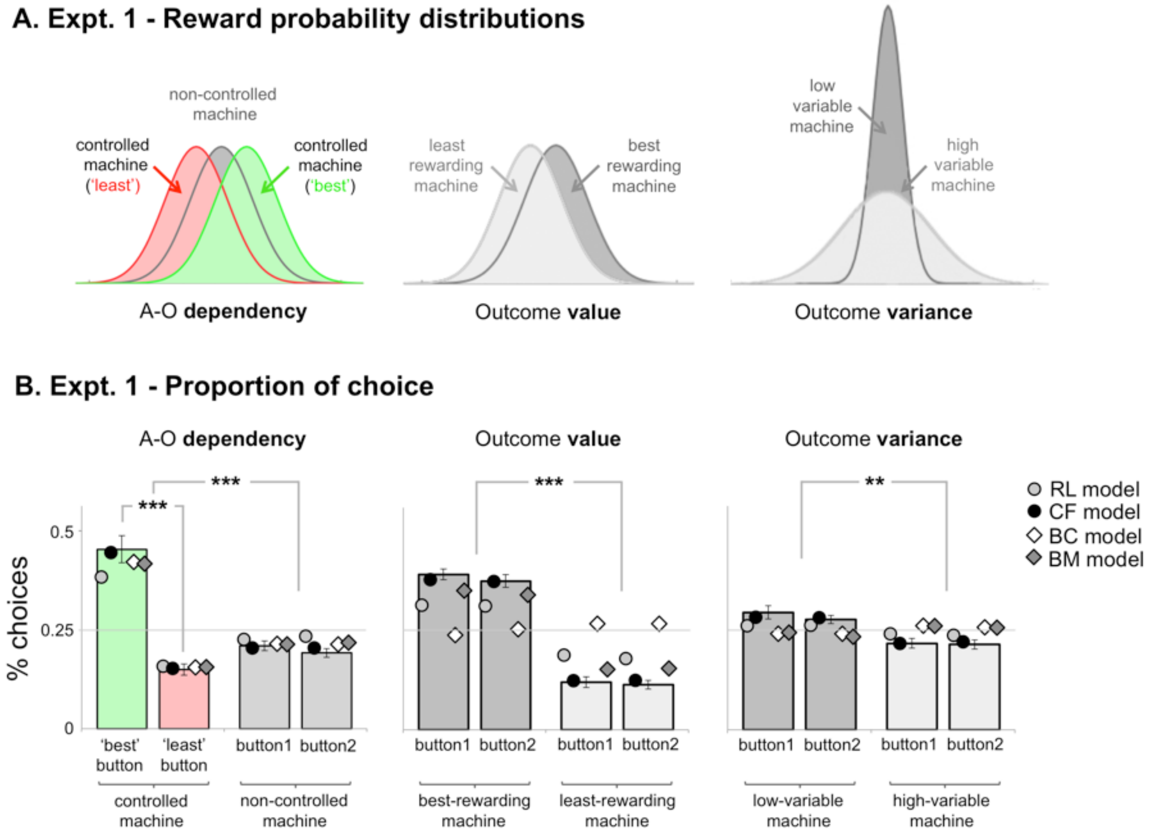


Figure 5: ***(A) Reward probability distributions associated with each button and machine of each experimental session from Experiment 1.*** Left panel: *In the first session, action-outcome dependency was implemented for one machine only. For this controlled machine, the outcome depended on the button choice: one button led to a mean outcome of 58 (green) whereas the other button led to a mean outcome of 42 (red). For the other uncontrolled machine (dark grey), the outcome displayed was drawn for a unique Gaussian distribution, irrespective of the button being chosen.* Middle panel: *In the "value" condition, no machine was controlled, but one bandit was best rewarded (dark grey, mean outcome: 58) than the other (light grey, mean outcome: 42).* Right panel: *In the "variance" condition, the mean outcome (50) was the same for both machines, irrespective of the button chosen, but outcomes from one machine were more variable than outcomes from the other machine (light grey, SD = 15, vs. dark grey, SD = 5).* ***(B) Mean proportion of choice for the three sessions, and for each button and machine.*** *Bars: participants' choices (%); dots and diamonds: models' choices (%). RL: reinforcement-learning model; CF: counterfactual model; BC: Bayesian-controller model; BM: Bayesian-maximizer model. The horizontal grey line indicates chance level (0.25%). All error bars indicate standard error. For the sake of visibility, models' error bars are not shown. Three-stars indicates $p < 0.001$.*

"variance" statistics on choice, the last two sessions did not implement any "divergent" machines. As a result, participants had no real control over the gains produced by the machines. The reason for this was twofold. First, it allowed for assessing whether choice behaviours modified in situations where one was told that events in the task were under one's own control but where no true control in fact existed – such as in classical settings implementing the so-called "illusion of control" (Stefan and David, 2013). Second, the procedure allowed for testing how best fitting models – i.e., models that best accounted for participants' choice under normal conditions – did perform in a situation of illusory control, and how well these models effectively accounted for the participant's data in this situation.

Finally, to keep all sessions as similar as possible, the same instructions were delivered across all three sessions. Thus, instructions in the "value" and the "variance" sessions were the same as those given in the "dependency" session, meaning that participants were not told they had no control over the machines in these conditions. All participants always started with the "dependency" session, implementing divergent and non-divergent machines, followed by the value and variance sessions in counterbalanced order across participants.

Reversals

Each session comprised 32 "episodes". An episode referred to an uninterrupted series of trials before a reversal occurred. The number of trials within an episode was on average 26 but varied between 14 and 38 (uniformly jittered) so as to make reversals as unpredictable as possible. In the "action-outcome dependency" session, two types of reversal could occur: either the buttons or the machines reversed, such that the controlled machine became the non-controlled machine or the best-rewarding button became the least-rewarding button. As for the value and variance sessions, only non-divergent machines were implemented, so that only "machine" reversals occurred: either the best-rewarding machine became the least-rewarding machine ("value" session) or the low-variable machine became the high-variable machine ("variance" session).

Modelling

To simulate participants' choices, we implemented the same four models that were previously described (RL, CF, BC and BM). In order to test if participants would adapt their strategy to the session, we fitted the models' parameters separately across the three different experimental sessions. As mentioned above, in both the "value" and "variance" sessions instructions were the same as those delivered in the "dependency" session: participants were not told they had no real control over the machines. This was explicitly accounted for by in the two generative models (BC and BM) through implementing the same latent states (i.e., generative distributions) as in the "dependency" session. Thus, our two generative models assumed there were a controlled and a non-controlled machine in all conditions.

Experiment 1: results

Percentage of choices

We first assessed whether subjects could discriminate between the two (divergent) states of the controlled, relative to the non-controlled, machine, by comparing choice proportion for each button of each machine, within each session. As expected, participants discriminated well between the two buttons of the controlled machine in the dependency session (best- and least-rewarding buttons: 0.45 vs. 0.15, $t_{15} = 6.3, p < 0.001$, Figure 5B, green vs. red bars, left panel), while choosing equally button 1 and button 2 of the non-controlled machines in all sessions (all $t_{15} < 1.62$, all $p > 0.12$; Figure 5B, grey bars). We then compared button preferences *across* all sessions. To do so, we subtracted choice proportion for one button from choice proportion for the other button within each preferred machine, and compared the difference across sessions using a one-way ANOVA (dependency vs. value vs. variance). The ANOVA confirmed that "button" preferences differed across the 3 sessions ($F_{2,45} = 28.98, p < 0.001, \eta_p^2 = 0.56$). Thus, participants discriminated between buttons of the preferred machine in the dependency session to a far greater extent than in the value and variance sessions (0.30 vs. 0.016, and 0.30 vs. 0.017, respectively, post hoc tests: all $p < 0.001$).

Second, we tested whether participants showed a preference for one machine over another within each session, by comparing choice proportion for each machine against the chance level (0.50). We found that participants showed a marked preference for the *controlled* machine in the dependency session ($t_{15} = 4.86, p < 0.001$), as well as a marked preference for the *best-rewarding* ($t_{15} = 10.67, p < 0.001$) and the *low-variable* ($t_{15} = 3.04, p = 0.004$) machines in the value and variance conditions, respectively. Finally, we compared the proportion of choice for the preferred machine across all 3 sessions. The one-way ANOVA revealed that "machine" preferences differed across the 3 sessions ($F_{2,45} = 21.31, p < 0.001, \eta_p^2 = 0.48$), Thus, participants chose the best-rewarding machine (value session, 0.76) more than the controlled machine (dependency session, 0.62), and both controlled and rewarding machines more than the low-variable machine (variance session, 0.57) (post hoc tests, all $p < 0.05$).

Note that, in all sessions, participants were able to quickly adjust to machine and/or button reversals: on average, the plateau of performance was reached within 5-10 trials after reversal (see Figure 6A, "Reversal learning curves").

Model comparison

Participants' trial-by-trial choice sequence were best accounted for by the CF model than by all other models in the set (RL, BM or BC). This was true for all conditions (exceedance probability > 98%) (Table 1 and Figure 6B). In addition to comparing model parameters across conditions and subjects, we also evaluated the generative performance of each concurrent model, i.e., its ability to replicate the participant's proportion of choices, but also the participant's trial-by-trial choice sequence after reversal (Palminteri et al., 2017). To do so, the 4 models were simulated with the best-fitting parameters on the whole experiment. Crucially, only the CF model showed a pattern of choices similar to that of participants in *all* sessions, whether with regard to the choice of the machine or to the choice of the button (see Figure 5B, CF = black circle).

Then, we plotted the models' learning dynamics before and after a reversal. Again, only the CF model was as flexible as participants, and adjusted to reversals with a similar dynamic (see Figure 6A, CF = red bars). In the dependency session, more specifically, the CF model outperformed all 3 competitors for both types of reversals. Thus, CF was able to retrieve the controlled machine and the best-rewarding option as quickly as participants, while the 3 other models adjusted more slowly, as particularly evidenced by the RL model after a button reversal (Figure 6A, top panel). In the value session, CF also better simulated participants' choices than all competing models (Figure 6A, bottom panel, left). Note that the BC model (dark green bars) aimed at maximizing control, i.e., preferentially chose the option associated with the controlled machine. Thus, its poor performance in this session with no true con-

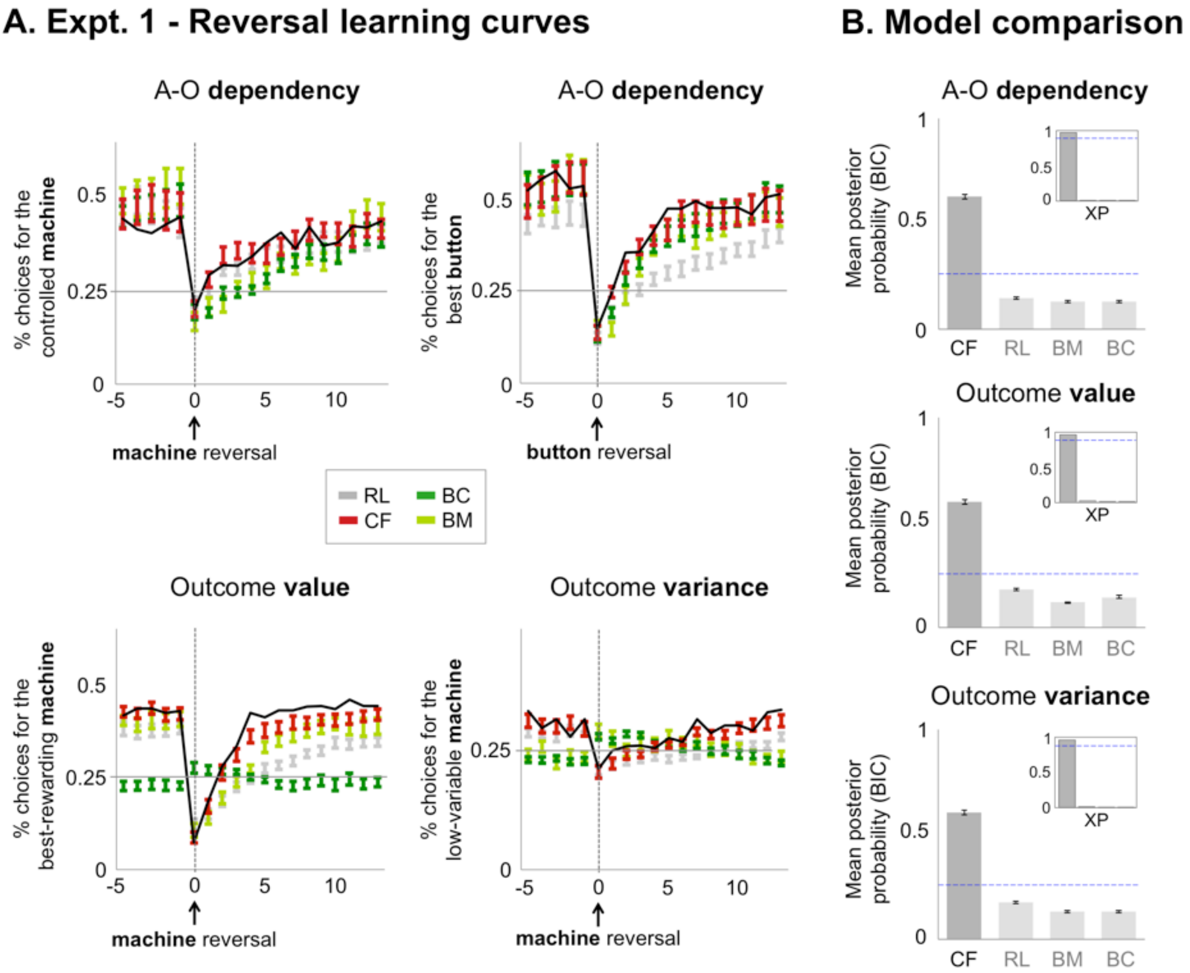


Figure 6: ***(A) Reversal curves for human participants (solid black line) and models (colored bars),*** *up to 15 trials after a "machine" or a "button" reversal.* Top panel: *reversal curves for the A-O dependency session, after a (controlled vs. non-controlled) machine reversal, or a (best vs. least-rewarding) button reversal.* Bottom panel: *reversal curves for the value and the variance session after a machine reversal (best vs. least-rewarding machine, or low vs. high variable machine, respectively). For the sake of readability, subjects' error bars are not shown. Model simulations: CF (red bars); RL (light grey); BM (light green bars); BC (dark green bars). Bars indicate standard error. RL: reinforcement-learning model; CF: counterfactual model; BC: Bayesian-controller model; BM: Bayesian-maximizer model. Dashed vertical lines indicate reversal point. Horizontal grey lines indicate chance level.* ***(B) Comparison of the posterior probability (PP) of each model, for each session.*** *The PP is calculated from the BIC, which penalizes model complexity. The blue dashed line represents the chance level at 0.25. The insert chart shows the exceedance probability (XP) of each model in the set. The blue vertical dashed line shows the 95% threshold. In all three sessions, the CF model best explained the data.*

trol comes at no surprise. The same remark applies to the variance session where no machine was controlled neither (Figure 6A, bottom, right). In this session, human participants showed a marked preference for the low-variable machine, and switched their choice after reversal so as to retrieve this preferred machine. Importantly, only the CF model was able to simulate this preference for poorly variable choice outcomes.

Experiment 1: discussion

The first experiment tested whether, and how well, human participants adjusted to self- vs. externally generated changes in a task where the source of these changes was uncertain.

Our results show that participants discriminated well between best- and least-rewarding buttons and between controlled and non-controlled machines. Hence, they preferentially chose the controlled machine over the non-controlled machine, while exhibiting a marked preference for both highly rewarding and low-variable machines. In the context of goal-directed control, this preference for high reward and low variance is reminiscent of the literature on self-attribution biases: adults are more likely to believe they control the occurrence of *positive*, relative to negative, events (e.g., Mezulis et al., 2004) while spontaneously assuming that series of *low-variable* events are more likely to be generated by intentional than non-intentional agents (e.g., Boland and

Table 1: *Mean (± s.e.m.) posterior probability (PP) of each model in each session and/or experiment. The exceedance probability (XP, in bold) refers to the probability that a given model fits the data better than all other models. CF: counterfactual model; RL: reinforcement-learning model; BC: Bayesian-controller model; BM: Bayesian-maximizer model.*

Expt.	Session	CF	RL	BM	BC
1a	Dependency	.60 (± .01) **99%**	.14 (± .005) **<1%**	.12 (± .003) **<1%**	.12 (± .003) **<1%**
	Value	.57 (± .01) **98%**	.17 (± .01) **<2%**	.11 (± .003) **<1%**	.14 (± .006) **<1%**
	Variance	.58 (± .01) **98%**	.16 (± .006) **<2%**	.12 (± .005) **<1%**	.12 (± .005) **<1%**
1b	Dependency	.47 (± .01) **92%**	.19 (± .006) **<4%**	.17 (± .005) **<3%**	.17 (± .005) **<3%**
	Value	.53 (± .01) **97%**	.18 (± .006) **<2%**	.14 (± .005) **<1%**	.13 (± .004) **<1%**
	Variance	.50 (± .01) **96%**	.17 (± .006) **<2%**	.16 (± .005) **<2%**	.15 (± .004) **<2%**
2		.85 (± .004) **99%**	.06 (± .002) **<1%**	.04 (± .001) **<1%**	.04 (± .001) **<1%**

Pawitan, 1999; Caruso, Waytz, and Epley, 2010). Unsurprisingly, the pattern of preference exhibited across all 3 sessions suggests that participants construe their action, not only as a mean to make a difference to the world (instrumental divergence), but also as an instrument to bring about positive events, and to reduce the inherent variability of the environment.

Both quantitative (BIC) and qualitative (simulated learning curves) results showed that a model drawing on pure associative processes (RL) cannot fully explain participants' behaviours, nor can generative models making choices based on either gain (BM) or control (BC) maximization strategies. Rather, we found that a model (CF) deriving the consequences of the forgone action from the current action taken, and assuming relative (i.e., context-dependent) divergence between both, best explained the data.

While BC and BM models had explicit priors about control in the task – assuming distinct outcome distributions depending on the subject's choice –, the CF model was endowed with a more general prior about instrumental divergence. This prior implements the belief that taking a specific action (e.g., choosing option A vs. B) makes a difference in terms of the outcome. Importantly, instrumental divergence is a reliable proxy for goal-directed control as the greater the action *diverges* with respect to its contingent states (the factual and counterfactual outcomes), the more flexible control one has over the environment. The fact that CF best explains data in all conditions suggests that human subjects construe their causal power based on such a prior. Interestingly, the CF model also best accounted for the participants' choice even when no true control existed, suggesting that this prior holds as a default belief, whereby goal-directed actions are thought to be causally efficient (i.e., divergent) in nature.

This study had two limitations. First, all sessions were not fully counterbalanced between subjects. All participants underwent the dependency session first, and then the two remaining sessions, where no true control was implemented. Besides, instructions given across all three sessions systematically emphasized the notion of control over the task. In a follow-up experiment (n=20), we thus ran a similar task while carefully controlling for these two potential biases. Sessions were fully counterbalanced and verbal and written instructions were kept as minimal as possible (see Appendix B, "Experiment 1b: method and results").

Most of the results from experiment 1 were replicated. As expected, participants exhibited a strong preference for high rewards and preferentially chose low-variable machines, regardless of their overall value. We also found that participants were able to discriminate between causally efficient actions, and to identify where in the task environment choosing one action rather than another made a difference to the outcome (the controlled machine), and where it did not (the non-controlled machine). Finally, we again found that a model based on a simple context-dependent counterfactual rule (CF) outperformed all competing models, including a pure reinforcement learner (RL) and a model that explicitly aimed at maximizing reward by means of Bayesian inference (BM).

In both experiment 1 and its follow-up, each statistic (dependency, value, variance) was tested within a different session, therefore limiting the opportunity to test and control for their potential interactions. In a second experiment, we addressed this limitation by implementing a factorial design where these statics were systematically crossed. In addition to controlling for interaction effects, this experimental design allowed for better characterizing subjects' choices in situations where these statistics were explicitly conflicting.

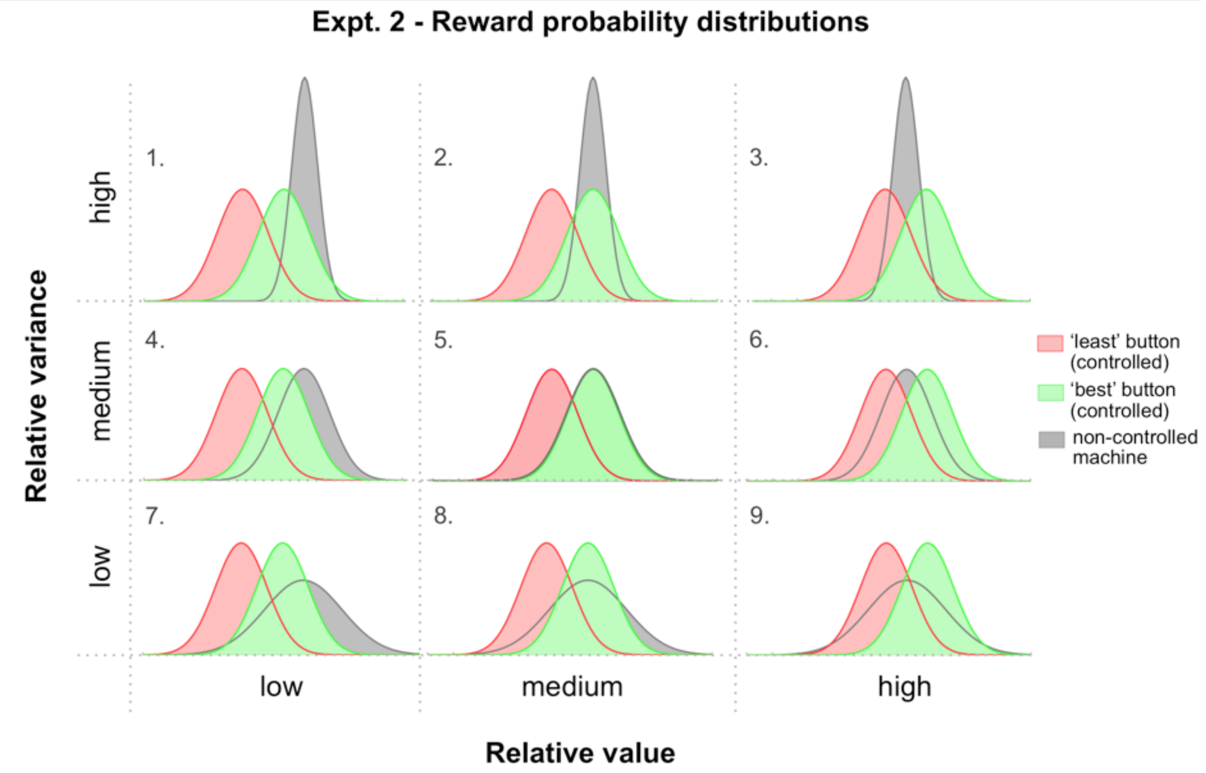


Figure 7: ***Schematic of experimental conditions in Experiment 2.*** *Contribution of action-outcome dependency (controlled vs. non-controlled machine), outcome value (x-axis) and outcome variance (y-axis), to the participant's choice, was assessed by manipulating the reward probability distributions associated with each button of each machine. Red and green Gaussian distributions: rewards from the controlled machine when the best- and least-rewarding buttons were selected, respectively. Grey distributions: rewards from the non-controlled machine, irrespective of the button selected.* X-axis: *the value of the controlled,* relative *to the non-controlled, machine, varied across three levels (low, medium, and high) – e.g., "low" level: the value of the controlled machine was low relative to the non-controlled machine.* Y-axis: *the variance of the controlled,* relative *to the non-controlled, machine, varied across three levels (low, medium, and high) – e.g., "low" level: the variance of the controlled machine was low relative to the non-controlled machine. Experimental conditions are numbered from 1 to 9 (top left to bottom right).*

Experiment 2: method

Participants

26 participants (14 females, age between 21 and 40 years-old) took part in Experiment 2. As before, they provided written informed consent prior to the experiment and were all paid 80 euros for the whole experiment (4 sessions). No participants had a history of neurological or psychiatric disorder, and all had a normal or corrected-to-normal vision. The experiment was approved by the local ethics review board (CCP C07-28). Participants were informed about the general procedure of the experiment through detailed written instructions.

Experimental sessions

The task (stimuli, timeline, and trial structure) was identical to that used in experiment 1. The only difference was implemented by the experimental design. Each participant completed 4 sessions, each lasting approximately 1 hour. Each session was carried out on a different day. A session consisted in 9 experimental conditions of 140 trials each (Figure 7), and was preceded by a brief training (64 trials). The order of conditions was pseudo-randomized within each session.

As in the previous experiment, the "statistical dependency" between action and outcome was manipulated by implementing a controlled (divergent) machine for which each button led to a different outcome, and a non-controlled (non-divergent) machine for which the reward was the same, regardless of the button pressed. For the controlled and non-controlled machines the gains associated with each button were discretized rewards drawn from Gaussian probability distributions, whose variance and mean depended on the condition (see below). The "value" and "variance" dimensions were crossed within a 3-by-3 factorial design, with each dimension varying across three levels (Figure 7):

1. The **value** dimension referred to the mean of the reward probability distribution associated with each machine. The mean of the controlled ma-

chine could vary across three different values (i.e., low = 42, medium = 54, high = 62) whereas the mean of the non-controlled machine was kept constant (i.e., 62). Low, medium, and high value characterized the average reward delivered by the controlled, relative to the non-controlled, machine (Figure 7, x-axis). The "low" value level indicated that that the controlled machine was on average less rewarding than the non-controlled machine, whereas the "high" level indicated that the non-controlled machine was more rewarding than the non-controlled one.

2. The **variance** dimension referred to the standard deviation (SD) of the reward probability distribution associated with each machine. The standard deviation of the controlled machine was kept constant all over the task (SD = 10) whereas the standard deviation of the non-controlled machine varied across three levels (low, SD = 5; medium, SD =10; high, SD = 15) (Figure 7, y-axis). These 3 levels characterized the variability of the rewards delivered by the controlled, relative to the non-controlled, machine. Thus, the "low" variance level indicated that the controlled machine was less variable than the non-controlled machine, whereas the "high" variance level indicated that the controlled machine was more variable than the non-controlled one.

Because the 3 levels of each dimension characterized the value and variance of the controlled machine *relative* to the non-controlled machine, we now refer to these as low, medium, and high, "relative levels". In the following, we looked at whether choice proportion changed as a function of the relative value and relative variance of the controlled machine. Specifically, we asked whether the proportion of choice for the best-rewarding button and/or the controlled machine would change as the controlled machine became more or less rewarding, or more or less variable, than the non-controlled machine.

Reversals

Finally, button or machine reversals could occur within each experimental condition as before. Button reversals consisted in the best-rewarding button (e.g., the square) becoming the least-rewarding button (e.g., the circle) of the controlled machine, whereas machine reversals consisted in the controlled machine becoming the non-controlled machine. Within each experimental condition, 6 reversals (3 machine reversals and 3 button reversals) could occur after a variable number of trials (between 14 and 26, uniformly jittered).

Experiment 2: results

Percentage of choices

We investigated the effect of the value and the variance dimensions, together with their interaction, on two dependent variables: (i) the proportion of choice for the controlled machine, and (ii) the proportion of choices for the best-rewarding button of the controlled machine. As in the previous experiment, the proportion of choice for the best-rewarding button was normalized by subtracting from it the proportion of choice for the least-rewarding button, within each condition.

The proportion of choices for the controlled machine, as well as the proportion of choices for the best-rewarding button, were analysed using two 3 × 3 repeated-measures ANOVAs, with the value (low vs. medium vs. high) and variance (low vs. medium vs. high) as within-subjects factors. Participants discriminated well between the controlled and non-controlled machines across all 9 conditions, but the proportion of choice for the controlled machine differed significantly as a function of the dimension manipulated. Thus, we found a significant main effect of the value ($F_{2,50} = 283.50, p < 0.001, \eta_p^2 = 0.91$) and a significant main effect of the variance ($F_{2,50} = 5.48, p = 0.007, \eta_p^2 = 0.18$) factor on the proportion of choice for the controlled machine. The proportion of choice for the controlled machine progressively increased as its relative value increased (low < medium < high, post hoc test, all $p < 0.001$), but also when its relative variance decreased (low vs. medium, p = 0.009; medium vs. high, $p = 0.04$). These results are consistent with the high-value and low-variance biases observed in experiment 1, wherein participants tended to preferentially select the machine with the highest value and the lowest variance (see Figure 5).

The value-by-variance interaction effect was also significant ($F_{4,100} = 6.66, p < 0.001, \eta_p^2 = 0.21$). Thus, when the relative value of the controlled machine was high, participants more often chose this machine, irrespective of the variance dimension; that is, they chose the controlled machine in similar proportions whether the controlled machine was highly or poorly variable (post hoc tests comparing low vs. medium vs. high variance, all $p > 0.12$). On the other hand, when the value of the controlled machine was low, participants tended to choose the controlled machine more when it was poorly, rather than highly, variable (comparing low vs. medium variance, $p = 0.07$; low vs. high variance, $p = 0.005$). In other words, the variance dimension had the strongest effect on the choice of the controlled machine when the value of this machine was the lowest (Figure 8, top panel, "Machine choice", "LOW value").

As in experiment 1, we then compared the proportion of choice for the best-rewarding button *across* conditions. We again found significant main effects of the value ($F_{2,50} = 78.81, p < 0.001, \eta_p^2 = 0.75$) and variance ($F_{2,50} = 4.28, p < 0.019, \eta_p^2 = 0.14$) dimensions.

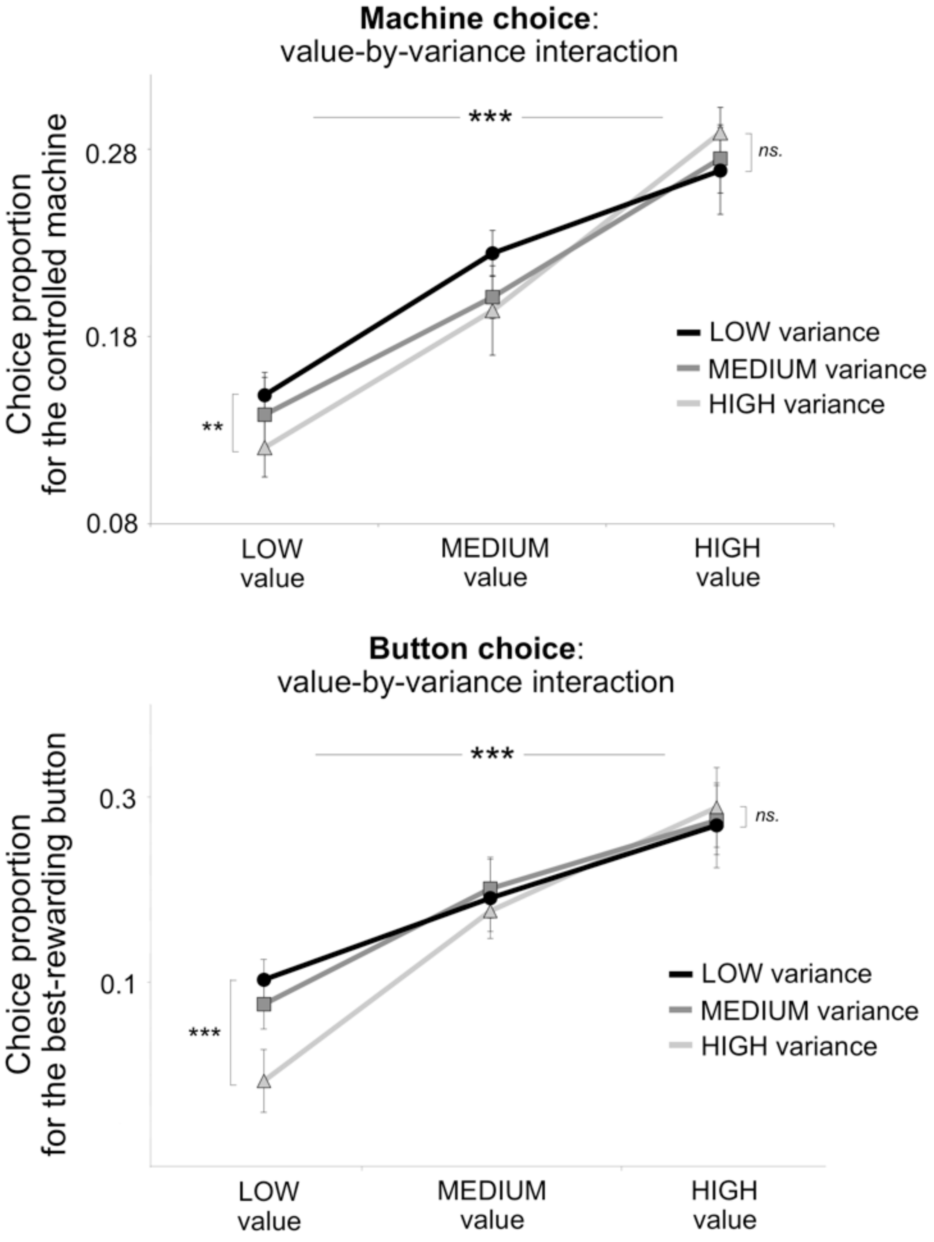


Figure 8: ***Participants' performance: mean proportion of choice (± s.e.m.) across each dimension manipulated.*** Top panel: *proportion of choice for the controlled machine in trials where the machine had a low, medium, or high value, relative to the non-controlled machine (x-axis), and had a low, medium, or high, variance, relative to the non-controlled machine (black, dark grey, and light grey, solid lines, respectively). The interaction effect between the value and variance factors was significant: the variance dimension had the strongest effect on the choice of the controlled machine when the value of this machine was the lowest.* Bottom panel: *normalized proportion of choice for the best-rewarding button in trials where the controlled machine had a low, medium, or high value, relative to the non-controlled machine, and had a low, medium, or high, variance, relative to the non-controlled machine. As for the choice of the controlled machine, the interaction effect between the value and variance factors was significant. Two-stars: $p < 0.005$; Three-stars: $p < 0.001$; ns. = $p > 0.05$.*

With respect to the value dimension, the higher the relative value of the controlled machine, the more often participants chose the best, relative to the least, rewarding button (post-hoc tests: low vs. medium= 0.28 vs. 0.17, $p < 0.001$; medium vs. high, $p < 0.001$). In other terms, the more rewarding was the controlled, relative to the non-controlled, machine, the more participants discriminated between each button, and the more their choice reflected the true "divergence" of the controlled machine (Figure S2). The same was observed for the variance dimension: the proportion of choice for the best, relative to the least, rewarding button increased as the relative variance of the controlled machine decreased (post hoc tests comparing low vs.

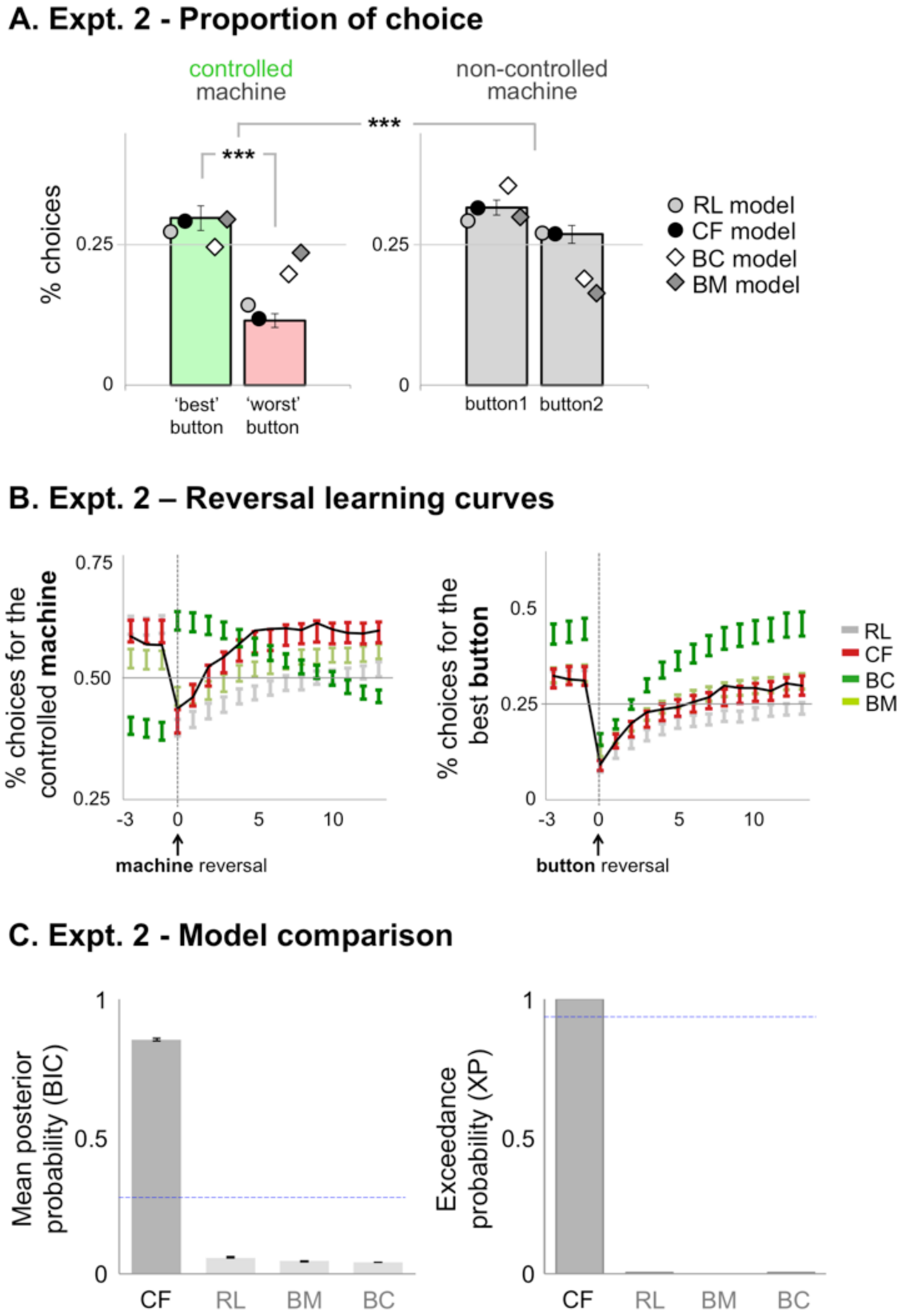


Figure 9: ***(A) Bar graphs comparing the proportion of choice (± error bars) across buttons and machines, averaged across all dimensions of the task design.*** *Bars: participants' performance; dots and diamonds: models' performance. For the sake of visibility, models' error bars are not shown. Three-stars: $p < 0.001$.* ***(B) Reversal curves (± error bars) for participants (solid black line) and models (colored bars)****, up to 15 trials after a button reversal. The horizontal grey line indicates chance level. Dashed vertical lines indicate reversal point (left graph: machine reversal; right graph: button reversal). Model simulation: CF (red bars); RL (light grey); BM (light green bars); BC (dark green bars).* ***(C) Comparison of the posterior probability (PP) of each model, for each session.*** *The PP is calculated from the BIC, which penalizes model complexity. The blue dashed line represents the chance level at 0.25. The right graph shows the exceedance probability (XP) of each model in the set, with the blue dashed line representing the 95% threshold.*

medium variance: 0.15 vs. 0.17, $p = 0.07$; medium vs. high variance: 0.17 vs. 0.19, $p = 0.005$) (Figure S3). Finally, the value-by-variance interaction effect was also significant ($F_{4,100} = 8.54, p < 0.001, \eta_p^2 = 0.25$). We found the same pattern of interaction as for the machine choice: the variance dimension had the strongest effect on button choice as the value of the controlled machine decreased (Figure 8, bottom panel, "Button

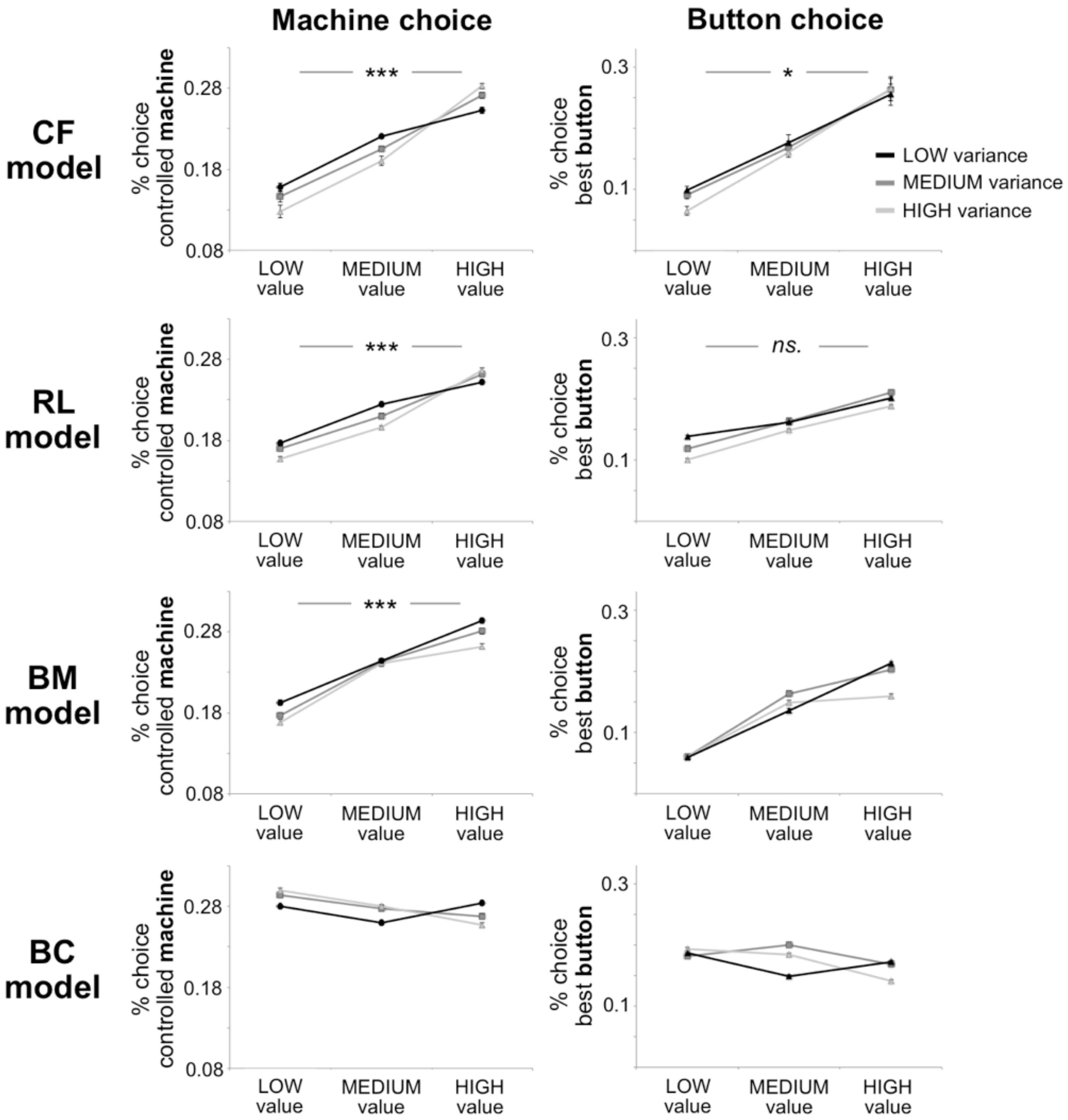


Figure 10: ***Models' performance: mean proportion of choice (± s.e.m.) across each dimension manipulated.*** *Stars indicate a significant interaction between the value and variance factors. Only the CF model replicate the interactions observed in human subjects, for both machine (left) and button (right) choices. The BC model actually shows the* opposite *interaction effects. One-star: $p < 0.05$; Three-stars: $p < 0.001$; ns. = $p > 0.05$.*

choice", "LOW value").

In sum, for both dependent variables (machine and button choices), the outcome value had an overwhelming influence on participants' choice, and this influence largely overrode the effect of variance. As a consequence, the effect of variance could only be observed in conditions where the value of the controlled machine was the lowest (see Figure 8, top and bottom panels).

Model comparison

The same four models were fitted and simulated to the data as before. Again, the CF model best predicted participants' choices (exceedance probability = 99%, Table 1, and Figure 9C), whether with regard to choice proportion for the best-rewarding button (Figure 9A, CF = black circles) or to choice proportion along the value (Figure S2) or the variance (Figure S3) dimensions.

Importantly, the participant's sensitivity to action-outcome dependency was best accounted for by the CF model. Thus, CF was the only model that did not underestimate the difference in choice proportion between buttons of the controlled and non-controlled machines (see Figure 9A, black circle). The CF model also correctly simulated the participant's choices along the value dimension. Thus, the CF model was able to reproduce the participant's propensity to better discriminate between the "best" and "worst" buttons as the value of the controlled, relative to the non-controlled, machine increased (Figure S2). The RL (grey circles) and BM model (grey diamonds) showed a similar, although less clear-cut, pattern of choice. In contrast, the BC model (white diamonds) exhibited the same pattern of choices across all 3 levels of the dimension, and both BC and BM models underestimated the difference between the two buttons of the controlled machine in the high value condition. The same applied to the variance dimension: both the CF and RL models were able to discriminate buttons of the controlled machine while choosing equally often the two buttons of the non-controlled one (Figure S3). In contrast to the BC and BM models, CF and RL also tended to more often choose the low-variable, relative to the high-variable, machine, as participants did.

As for participants, models' choices for the controlled machine and for the best-rewarding button were analysed further using two 3 × 3 ANOVAs, with value (low vs. medium vs. high) and variance (low vs. medium vs. high) as within-subjects factors. Relative to participants' performance, only the CF model was able to replicate the main effects of the value (machine choice: $F_{2,50} = 289.36, p < 0.001, \eta_p^2 = 0.92$; button choice: $F_{2,50} = 86.00, p < 0.001, \eta_p^2 = 0.77$) and of the variance (machine choice: $F_{2,50} = 2.50, p < 0.001, \eta_p^2 = 0.33$; button choice: $F_{2,50} = 5.03, p = 0.01, \eta_p^2 = 0.16$) factors, as well as the significant interaction effects between them (machine choice: $F_{4,100} = 21.98, p < 0.001, \eta_p^2 = 0.46$; button choice: $F_{4,100} = 2.68, p = 0.035, \eta_p^2 = 0.09$; post hoc comparing low vs. high variance, $p = 0.032$) (Figure 10, "CF model"). Importantly, none of the 3 other models could replicate this exact pattern of performance.

The CF model also showed the most consistent reversal curves across conditions, outperforming all competing models when adjusting to changes according to either action-outcome dependency (Figure 9B), outcome value (Figure S2) or outcome variance (Figure S3). Specifically, both the BM and CF models correctly simulated the participant's learning curves, whether in terms of dynamic (slope) and absolute performance (plateau), while the RL model converged to the plateau of performance more slowly than real subjects did. Note that the BC model (dark green line) was designed to preferentially choose the action associated with the controlled machine, and hence systematically reversed choice after reversal of the best button.

Experiment 2: discussion

Experiment 2 reproduced most of the effects previously obtained, in a design controlling for potential interaction effects between conditions. Importantly, participants performed the task well despite no explicit cue was available to signal the transition from one condition to the other. In a situation where uncertainty was high, participants were able to monitor the different statistics implemented by the task, and to adjust when these statistics changed and reversals in (either machine or button) contingencies occurred. In line with experiment 1, we found that participants chose more often the *controlled* machine when the relative *value* of this machine increased, but also when its relative *variance* decreased, consistently with the literature on self-attribution biases. Likewise, when the value of control increased, participants discriminated better between the best and worst option of the controlled machine, and choice behaviour was hence found to better reflect the true divergence of the controlled machine. Finally, we found a significant interaction between value and variance factors. Specifically, a significant effect of *variance* on machine and button choice was observed in *low-value* trials only. This interaction suggests that competition between both statistics is fundamentally asymmetrical. In case of a conflict, subjective preferences for highly valued options overrode preferences for options giving rise to poorly variable outcomes. On the other hand, when the difference in value between competing options was low, subjects made a choice based on variance estimates from past choice outcomes.

As expected, the overall effect of value on choice was well captured by algorithms that aimed at maximizing rewarding value (RL, CF), while an optimal learner aiming to maximize control (BC) failed to account for this effect. We again found that the CF model outperformed all competitors according to both quantitative (BIC) and qualitative (reversal curves) criteria. Interestingly, CF was the only model to not systematically underestimate the difference in choice proportion between the two buttons of the controlled and non-controlled machines, and to better discriminate between each button of the controlled machine as the value of this machine increased, as participants did. The CF model was also the only model to replicate the exact pattern of performance found in human subjects. Thus, the CF model increasingly chose the controlled machine and the best-rewarding button as the value of this machine increased (main effect of value), but also as its variance decreased (main effect of variance). Critically, choices of the CF model also exhibited a significant value-by-variance interaction effect. Thus, the effect of variance on the model's choices was only observed in low-value trials, as in human participants. In sum, we found that one single model (CF) was able to simulate participants' performance across all three dimensions, and was able to do so with the same set of parameters and same parameter values.

In the next section of the paper, we analysed and compared further the parameters of the "winning" CF model across both experiments 1 and 2, namely: (1) the reference point, and (2) the factual and counterfactual learning rates. Importantly, these two sets of parameters can be seen as direct or indirect proxies for instrumental control:

1. The "reference" is a fitted parameter whose value approximated the mean of the reward distribution associated with the chosen action (see Figure 11A and 11B). It is an add-on to the classical RL algorithm, that implements control as difference-making. Thus, the more the value of the reference departs from the true reference, the more divergent actions are, that is the greater the *difference* between the outcomes associated with the chosen and unchosen actions (see Figure 11C, for an illustration).
2. The "counterfactual (CF) learning rate" is a proxy for how much weight is given to the counterfactual prediction error. In a world where instrumental control is assumed (i.e., a world where factual and counterfactual actions give rise to different outcomes), a CF learning rate is a measure of how fast

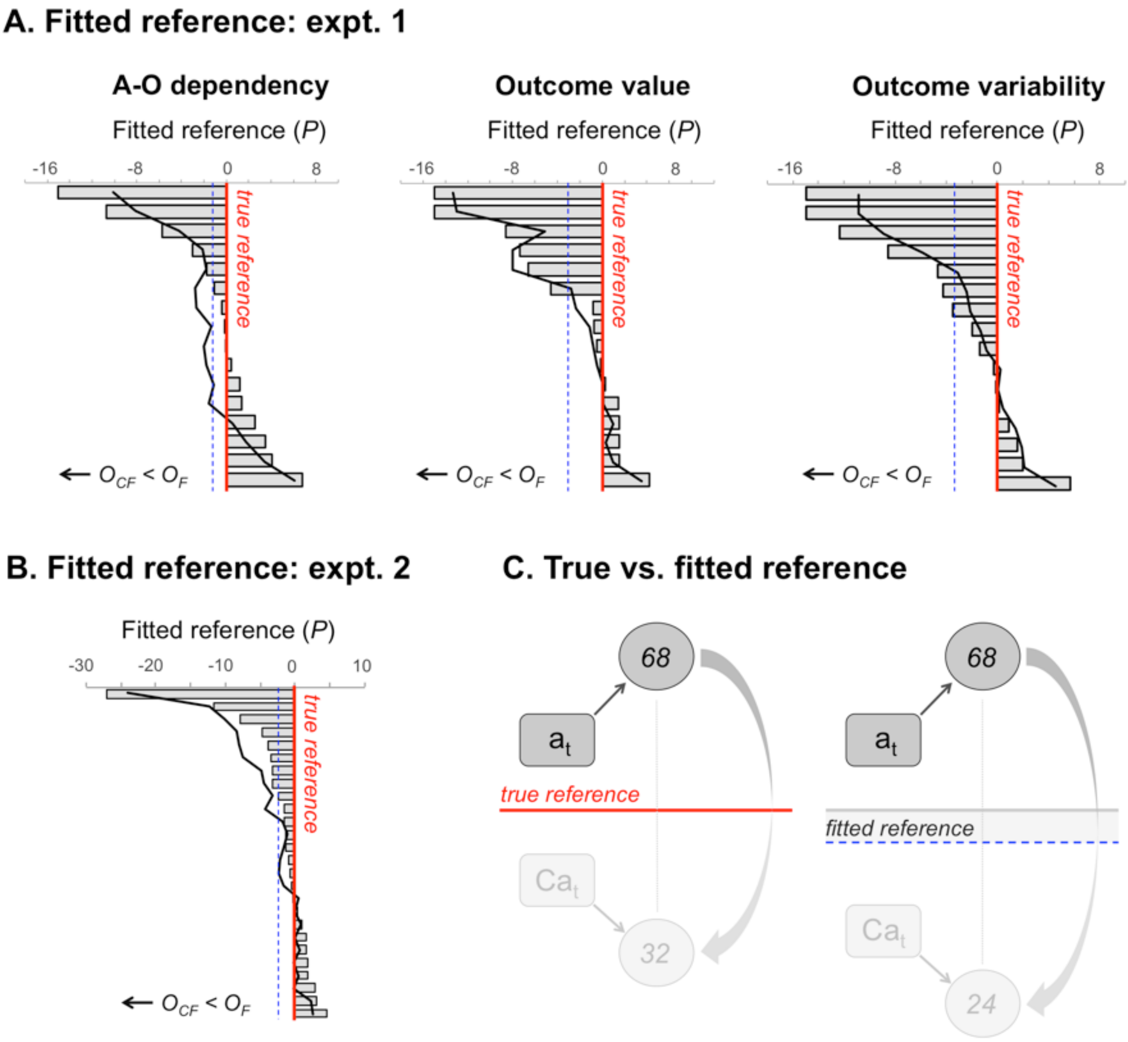


Figure 11: ***Fitted individual references across the different sessions of Experiment 1 (A) and Experiment 2 (B).*** *The bars represent the value of the fitted reference* relative *to the true reference, i.e., the true mean of the reward distributions (vertical red line), in each participant. A negative value indicates that the participant underestimated the true reference. The greater the negative value, the lower the counterfactual reward inferred by the subject, relative to the factual reward ($O_{CF} < O_F$). Conversely, the greater the positive value, the greater the counterfactual reward inferred by the subject, relative to the factual one. Below the red line, the vertical dashed blue line represents the group mean of the fitted reference. Over individual bars, the solid dark curve represents the* divergence *between chosen and unchosen alternatives in each subject. The divergence was calculated by subtracting the factual from the counterfactual reward on each trial, based on the subject's fitted reference, and averaging the result over all trials.* ***(C) True vs. fitted reference.*** *When combined with the contextual rule of the CF model, underestimating the true reference leads to exaggerating the divergence between factual and counterfactual outcomes (e.g., 24 rather than 32).*

the *divergence* between factual and counterfactual outcomes builds up over time.

CF model: best-fitting parameters

We first compared the value of the reference parameter against the "true" reference, i.e., the true mean of the reward distributions, in both experiments. The value of the fitted reference overall approximated the mean of the reward distributions (t-tests against the mean of the reward distributions in each experiment: all $p > 0.05$, except for the variance condition: $p = 0.04$, see Figure 11). Note that the value of the fitted reference varied across subjects, with some participants substantially underestimating the true mean of the current distribution (see Figure 11, vertical dashed black lines). Interestingly, participants who underestimated the true mean also tended to exaggerate instrumental divergence as a result – i.e., the difference between chosen (factual reward) and unchosen (counterfactual reward) alternatives (see Figure 11A and 11B, dark solid curve, and 11C).

We next compared factual and counterfactual learning rates within and between experiments. Three counterfactual alternatives were updated on each trial:

- the unchosen button of the chosen machine (α_{CF1}),
- the chosen button of the unchosen machine (α_{CF2}),
- the unchosen button of the unchosen machine (α_{CF3}).

To first compare the factual and counterfactual learning rates of experiment 1, we carried out a 2 × 2 × 3 repeated-measures ANOVA, with the button (chosen vs. unchosen), the machine (chosen vs. unchosen), and the 3 different statistics (dependency vs. value vs. variance), as within-subjects factors. A similar 2 × 2 repeated-measures ANOVA was performed on all pooled conditions of experiment 2.

In experiment 1, the main effect of the "machine" ($F_{1,15} = 4.66, p < 0.03, \eta_p^2 = 0.273$) and the main effect of the "button" ($F_{1,15} = 15.78, p = 0.005, \eta_p^2 = 0.51$) were significant. Thus, the learning rate associated with the chosen machine was significantly lower than the learning rate associated with the unchosen machine (post-hoc test, all $p < 0.04$), whereas the learning rate associated with the chosen button was globally higher than that of the unchosen button (all $p < 0.001$).

The machine-by-button interaction effect was also significant ($F_{1,15} = 60.07, p < 0.001, \eta_p^2 = 0.80$). Across all three sessions of experiment 1, post hoc tests showed that learning rates of the chosen buttons did not differ across chosen (α_F) and unchosen (α_{CF2}) machines, while learning rates associated with the unchosen buttons (α_{CF1} vs. α_{CF3}) differed significantly (all $p < 0.001$). Chosen (α_F) and unchosen (α_{CF1}) buttons of the chosen machine also differed significantly (all $p < 0.001$), while they could not be distinguished for the unchosen machine (α_{CF2} vs. α_{CF3}) (Figure 12A). This interaction effect was observed in all conditions equally (i.e., no significant modulation of the machine-by-button interaction by the type of statistics: $F_{1,15} = 1.17, p = 0.3, \eta_p^2 = 0.07$). Experiment 2 showed the same tendency as in the previous experiment (see Figure 12B). However, only the machine-by-button interaction effect was statistically significant ($F_{1,25} = 4.06, p = 0.04, \eta_p^2 = 0.21$). Note that the pattern of fitted learning rates from the CF model was correctly recovered when applying the procedure to simulated data, and hence was not an artefact of the parameter optimization procedure (see Figure 12C, and Appendix D, "Parameter recovery procedure").

Interestingly, our findings reveal that participants calibrate their learning rates in a way that reflects their belief about the task structure. First, counterfactual (CF) learning rates associated to the button or to the machine were significantly higher than zero in all experiments and conditions (see Figure 12, comparing α_{CF1}, α_{CF2}, $\alpha_{CF3} > 0$, all $p < 0.005$). A high CF learning rate indicates that participants update the value of the forgone alternative; this counterfactual update results in making the value of the unchosen alternative diverge from the value of what is currently chosen. Thus, above-zero CF learning rates show that our participants construed their actual choice as being *causally efficient*, i.e., as making a difference relative to the unchosen alternative.

A CF learning rate is formally equivalent to the notion of "mutability" in previous work on counterfactual reasoning (e.g., Dehghani, Iliev, and Kaufmann, 2012; Kahneman and Miller, 1986). Mutability is a property of a variable that signals whether the variable is likely to take different values in the real and counterfactual worlds. Thus, a *highly mutable* variable is highly likely to diverge across factual and counterfactual worlds (Lucas and Kemp, 2015). Similarly, a machine associated with a *high CF learning rate* is a highly mutable machine: choosing this machine, rather than the other one, should make a significant difference with respect to the outcome. Conversely, a low CF learning rate would minimize the divergence, while a *null* CF learning rate would signal a *null* divergence between the chosen and unchosen options. Importantly, our results revealed a hierarchy across buttons and machines. Counterfactual learning rates were higher for the *machine* than for the button, suggesting that participants conceived the former as being more "mutable" than the latter. In other words, participants considered that making a choice about the machine was more likely to make a difference to the world relative to making a choice about the button.

What does this hierarchy account for? We suggest that counterfactual emulation is more likely to be leveraged for testing control at most abstract levels of action

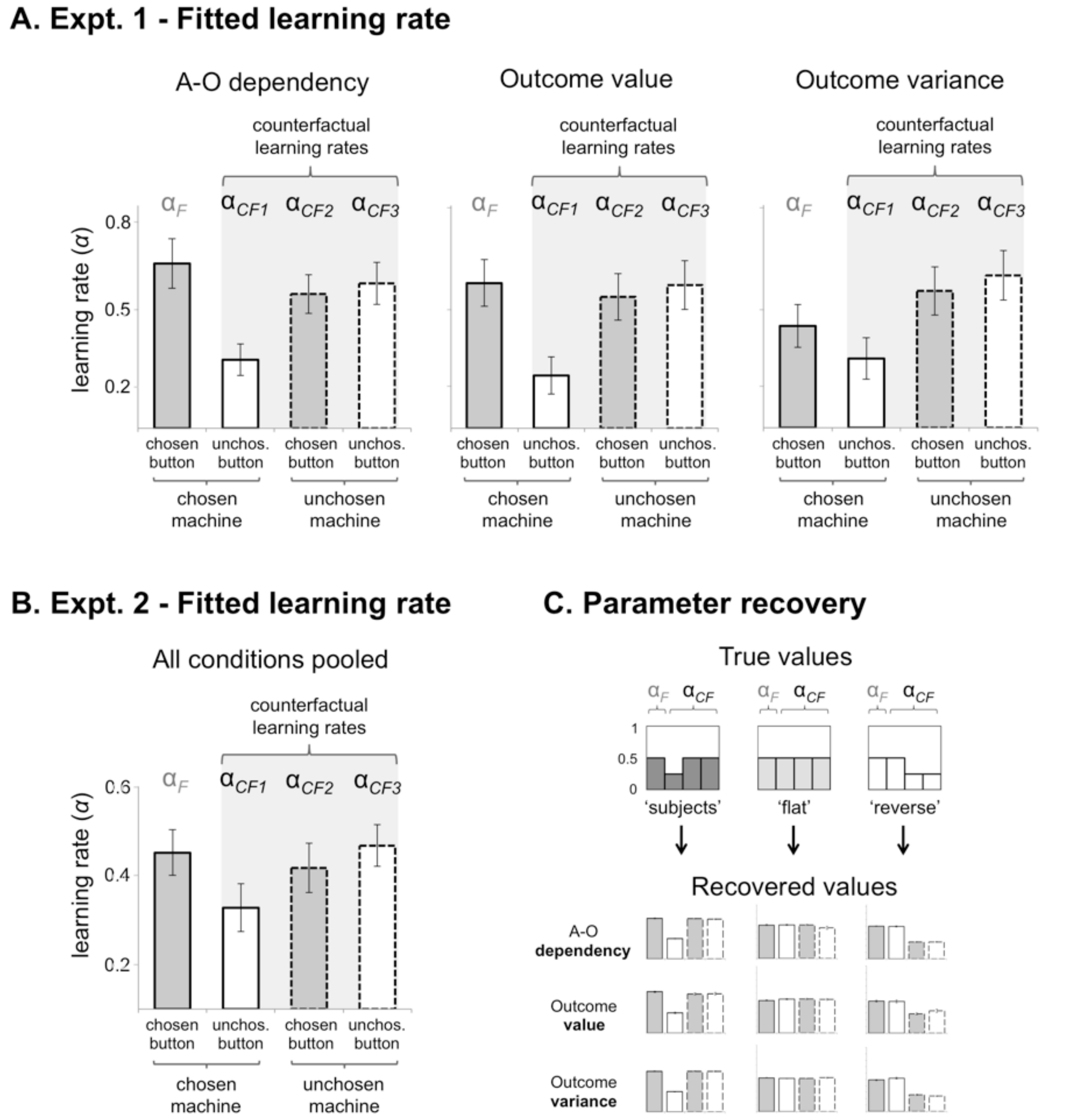


Figure 12: ***Fitted learning rates from the winning model (CF). (A) Expt. 1***: *Factual (α_F) and counterfactual learning rates (α_{CF}) within each experimental session, for each button (chosen, unchosen) and each machine (chosen, unchosen).* ***(B) Expt. 2***: *Factual and counterfactual learning rates for all conditions pooled together.* ***(C) Parameter recovery procedure***: *"True value": learning rates used to simulate the data (see Appendix D, Table S1). "Recovered value": learning rates obtained from fitting the model on the simulated data. "Subjects" = highest learning rate for the unchosen button; "flat" = identical learning rates across the unchosen button and the unchosen machine; "reverse" = highest learning rates for the unchosen machine. Our parameter optimization procedure was able to correctly recover the (true) parameter values from all patterns in all sessions.*

representation (e.g., at the level of the machines), and less required for less abstract levels (e.g., the level of the buttons) where direct instrumental testing is available to the subject. Crucially, should this prediction be correct, counterfactual emulation would be optimal in an environment where instrumental divergence is maximal between *machines*, rather than between buttons.

We directly tested this hypothesis by simulating our CF model across two different environments: (1) an environment where divergence was maximal between buttons, or (2) an environment where divergence was maximal between machines (see Figure 13A, left and right panels, respectively, and Appendix D "Performance of the CF model"). We tested the performance of different patterns of learning rates across these two types

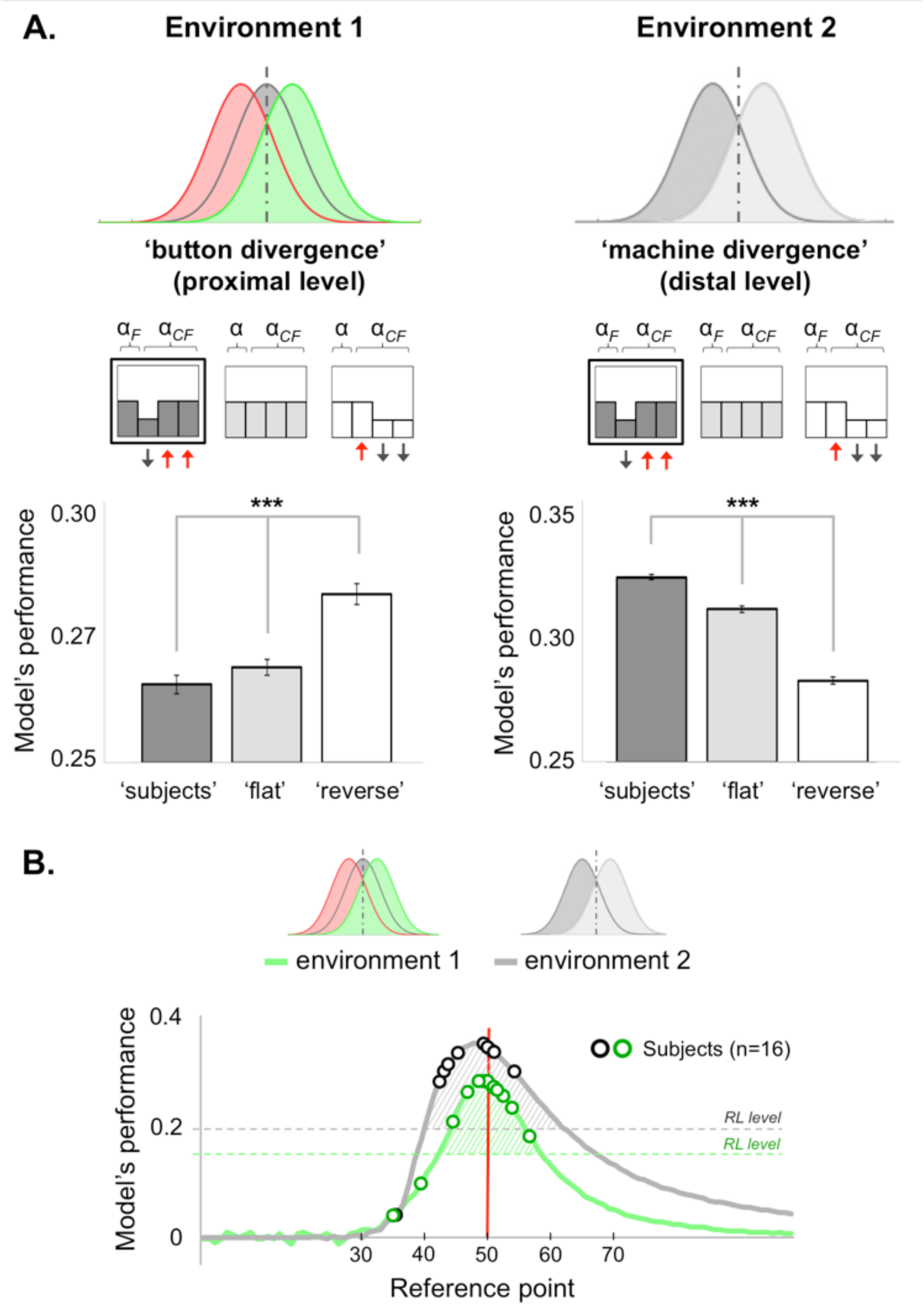


Figure 13: ***(A) Simulated performance of the winning model (CF) across two different environments, with three distinct patterns of learning rate.*** Left panel: *performance of the CF model when the divergence is maximal between buttons (red and green distributions). The pattern of participants (dark box: "subjects") is outperformed by the two alternative patterns ("flat" and "reverse").* Right panel: *performance of the CF model when the divergence is maximal between machines (dark and light grey distributions). Here, the pattern of participants outperforms the two alternative patterns. Model's performance is normalized against chance-level.* α_F *= factual learning rate (chosen button of the chosen machine);* α_{CF} *= counterfactual learning rate (unchosen button and/or unchosen machine). Three-stars:* $p < 0.001$. ***(B) Averaged performance of the CF winning model (y-axis) across the two environments, depending on the value of the reference point (x-axis).*** *Model's performance is optimal in both environments when the reference point approximates the true mean (red vertical line), as most subjects did (green and dark circles). The hatched areas delineate the range of reference values for which the CF model outperforms the RL model (horizontal dashed lines). In both environments, model's performance is normalized against chance-level.*

of environment: i) a pattern that was similar to that of participants (*higher* CF learning rates for the most abstract level, i.e., the unchosen machine), ii) the reverse pattern (*lower* CF learning rates for the unchosen machine), and iii) a flat pattern (*equal* learning rates across unchosen machines and buttons) (see Appendix D, Table S1). Consistently with our prediction, we found that the pattern of participants outperformed the two alternative patterns in the environment where the divergence was set at the most abstract level, i.e., the machine level (see Figure 13A). Importantly, this was all the more true as values of the reference point approximated the true mean of outcome distributions (see Figure 13B).

General discussion

Using a modified reversal-learning procedure, we tested whether, and how well, human participants could adjust to self- vs. externally-generated changes in a task where the source of these changes was uncertain. Specifically, any perceived changes could potentially be ascribed to three different causes: (i) the participant's choice, (ii) the intrinsic variability of the outcome, or (ii) a reversal in either instrumental or environmental contingencies. Thus, maximizing performance in the task required the ability to discriminate action-related changes from changes due to intrinsic feedback noise and/or external volatility, and to adjust one's choice behaviours accordingly.

Behavioural results: control, value and variance

In all experiments, we found that participants were able to discriminate best- from least- rewarding buttons, and to distinguish between the controlled and non-controlled machines – that is, between the machine for which there was a best- and a least-rewarding button and the machine for which both buttons were equally rewarding. Participants performed the task well despite no explicit cue was available to signal reversals in the best-rewarding button or in the controlled machine. In experiment 1, participants preferentially chose the controlled machine over the non-controlled machine, while also exhibiting a marked preference for the best-valued and low-variable machines in both experiments 1 and 2. Interestingly, both outcome value and variance had an effect on the proportion of the controlled machine. Thus, participants more often chose the controlled machine when the relative value of this machine increased but also when its relative variance decreased. Likewise, when the value of control increased, participants discriminated better between the best and worst option of the controlled machine, and choice behaviour was hence found to better reflect the true divergence of the controlled machine.

This interaction between value and control, and between variance and control, is reminiscent of self-attribution biases, whereby healthy adults take credit for positive outcomes while denying responsibility for negative events (e.g., Mezulis et al., 2004), and overestimate the variability of random series while underestimating the variability of self-caused events (e.g., Boland and Pawitan, 1999). Spurious *positive* relationships between control and value are further exemplified in situations where people mistake the value of an event for real control over this event, through inflating probabilistic estimates of action-event contingencies (Kool, Getz, and Botvinick, 2013). This interplay between control, (high) value, and (low) variance, suggests that individuals construe the effects of their action along multiple dimensions: as a mean to make a difference to the world, but also as an instrument to bring about positive events, and to reduce the inherent variability of the environment. Importantly, we found that one single model (CF) was able to simulate participants' performance across all three dimensions, and was able to do so with both the same set of parameters and same parameter values.

Associative learning and counterfactual update

In both experiments, optimal performance required a complete knowledge of the underlying causal structure of the task, namely, a representation of each probability distribution relating each possible action to a particular state. Thus reaching optimal performance was computationally costly, as it ideally required maintaining probability distributions across all alternative causes and updating all possible alternatives at once, whenever integrating new evidence. Whether such a strategy is used, or even usable, by human subjects remains conjecture (Eckstein et al., 2004; Jones and Love, 2011). Although they lack an explicit representation of instrumental contingencies, simpler learning schemes, e.g., based on pure associative processes, can readily adapt to causally structured environments, at much a lesser cost (Dickinson, 2001). On the other hand, associative processes only enable a form of *proximal* instrumentality, whereby acquisition and performance of new and existing behavioural strategies are regulated by their immediate consequences. Accordingly, associative agents only slowly adapt to environments with periodically changing action-outcome mappings, and hence would hardly approximate the efficiency of human performance (Gershman, Markman, and Otto, 2014). An intermediate solution would consist in combining a (simple) associative learning scheme with a generative rule for emulating an approximate version of the environment's causal structure. In contrast with pure associative algorithms, this "combined" model would assume a generative source behind observation, but this source would not have to be a fully specified probability distribution of expected action outcomes. Models of *counterfactual reasoning* (e.g., Lucas and Kemp, 2015) can be specified so as to permit action outcomes to take different values in the real and counterfactual worlds. Importantly, these models can also account for hierarchical inference in causal reasoning by allowing factual and counterfactual action values to be updated according to different dynamics (e.g., learning rates).

We tested and compared the ability of associative, generative, and counterfactual models to account for the participants' data across all experiments. We found that models that merely aimed at maximizing action value – whether by prediction-error minimization (RL) or by means of Bayesian inference (BM) – could not explain the participants' choices well, neither could a model (BC) that aimed at maximizing control over the task, regardless of action value. On both quantitative (BIC) and qualitative (reversal curves) criteria, participants' behaviour was best accounted for by a model that made choices based on counterfactual contingencies, i.e., a model that emulated *unseen* action-outcome pairs according to a contextual rule. Thus, counterfactual contingencies were emulated by deriving the consequences of the forgone action from the current action taken and by assuming relative (i.e., context-dependent) *divergence* between both. Importantly, instrumental divergence was implemented in the model as a prior.

Specifically, this prior conveys the belief that taking a specific action (e.g., choosing option A vs. B) makes a difference in terms of the outcome. As mentioned above, instrumental divergence is a reliable proxy for goal-directed control as the greater the action *diverges* with respect to its contingent states (the factual and counterfactual outcomes), the more flexible control one has over the environment. The fact that the CF model best explained the data in all conditions suggests that human subjects construe their causal power based on such a prior. Interestingly, the CF model also best accounted for the participants' choice even when no true control existed (i.e., "value" and "variance" conditions), suggesting that this prior holds as a default belief, whereby goal-directed actions are thought to be *causally* efficient (i.e., divergent) in nature. Finally, only the CF model was able to replicate the value-by-variance interaction observed in subjects, for both machine and button choices, while the other models replicated this pattern only partially (e.g., RL) or exhibited the reverse pattern of performance (e.g., BC) (see Figure 10).

The counterfactual world negatively covaries with the real world

Counterfactual reasoning has been the subject of many investigations in the decision-making domain, from regret-based theory of choice (e.g., Coricelli et al., 2005; Bell, 1982) to empirical works on fictive learning, i.e., learning from alternative action values (e.g., Lohrenz et al., 2007). While it has been repeatedly shown that instrumental learning benefits from tracking alternative courses of action and their outcomes, how these counterfactuals are generated, and based on what rule, is currently unclear. In most studies on fictive learning, subjects are explicitly informed about the result of the forgone alternative (e.g., Lohrenz et al., 2007; Palminteri et al., 2015; Boorman, Behrens, and Rushworth, 2011). In our task, however, the reward associated with the unchosen machine was not shown to the participant but had to be inferred given the chosen button. Crucially, our CF model provides an algorithmic explanation for how counterfactual action values were inferred, based on a flexible, context-dependent, reference, whose value was fitted separately to each participant's data.

As previously argued, exploiting counterfactual information can be beneficial to the learner, provided the cost of getting and storing the information is not too high (Boorman, Behrens, and Rushworth, 2011). Importantly, the context-dependent reference embedded in the CF model approximated the mean of the generative distributions in the task, and thus allowed for emulating counterfactuals at low cost. The mean is a simple and often-optimal operator that affords minimizing error in error-prone situations. Under uncertainty, making decisions based on an averaged representation of the environment is often advantageous (Sutton and Barto, 1998). In this respect, the CF model would be efficient, not because it would maintain an expensive, yet accurate, causal model of the task (e.g., the probability distributions over all possibilities), but because it embeds a prior (the reference point) that approximates the actual structure of the environment (see also Parpart et al., 2017). In addition to showing better performance than a classical RL, the CF model also keeps simplicity in terms of algorithmic design and computation. We argue that this simplicity provides a step towards an explanation of how human agents achieve a trade-off between robust causal inference and the costs of maintaining an accurate model of the world (Bramley et al., 2017).

Updating the counterfactual according to a context-dependent reference is consistent with a broader literature on reference dependence in behavioural economics, where the utilities of outcomes are assessed relative to a context-specific reference point (e.g., Tversky and Kahneman, 1974; Kőszegi and Rabin, 2006; Denrell, 2015). Converging evidence from average-learning algorithms and computational neuroscience equally suggests that people make decisions according to a context-dependent reference (Palminteri et al., 2015; Klein, Ullsperger, and Jocham, 2017; Burke et al., 2016). Importantly, providing counterfactual information to the subject reinforces the dependence on context for evaluating rewards and punishments (Palminteri et al., 2015). Thus, when subjects are informed about the result of the forgone alternative, value contextualization is enhanced. Similar to our CF model, such contextualization would consist in tracking the mean of the distribution of values of the current choice context (i.e., the reference point), and using it to center both factual and counterfactual option values. Such value contextualization echoes adaptive coding of outcomes in neural populations, whereby neural outputs rescale to the range of currently expected outcomes (Burke et al., 2016), and more generally is consistent with studies on context-based processing of outcome information showing that motivationally relevant information is encoded in a relative fashion, adapting to the current value-context (Seymour and McClure, 2008).

Because it updates alternative action values based on a context-dependent reference, the CF model can be viewed as a generalization of the Rescorla-Wagner (R-W) rule (Sutton and Barto, 1998). Interestingly, counterfactual updating in associative learning can also be modelled using a Bayesian generalization of the R-W model, i.e., using Kalman filters based on temporal difference (TD) learning (Keramati, Dezfouli, and Piray, 2011). Kalman TD incorporates a component of counterfactual thinking by encoding a *negative covariance* between stimuli elements. In terms of instrumental learning, this covariance structure can be leveraged to update both factual and counterfactual action values, as learning one particular instrumental contingency automatically leads to a reduction in the associative strength of the unchosen contingency (Gershman, 2015). In a recent study, Morris et al. (2017) showed that instrumental learning was best explained by a Kalman algorithm that combines prediction-error learning with a similar covariance matrix, reflecting the structure of the task environment. In their task, several causal variables compete to explain the observation. The winning model assumes negative covariance between these variables, meaning that a change in the belief of one cause inversely affects the other (Morris et al., 2017). The covariance matrix thus allows the learner to reason *counterfactually* about alternative courses of action, hence to differentiate the unique effects of action from background effects, i.e., from effects that would have occurred in the absence of that action. Morris and collaborators found that this model, combining key features of associative learning and model-based RL, better characterized behavioral performance and neural activity associated with instrumental learning than models based on covariance or prediction-error alone.

Morris et al. (2017)'s model has formal resemblance with our CF model. In the CF model, however, the negative covariation between factual and counterfactual values critically relied on a parameter, the reference point, whose value was separately fitted in each participant. Importantly, the value of this reference showed some variability across participants, depending on their subjective preferences and beliefs. Thus, while some underestimated, some other overestimated, the true mean of the current distributions. Interestingly, underestimating the true mean was *self-serving* in nature, as it led to exaggerate the divergence between factual and counterfactual outcomes. Thus, in subjects underestimating the true mean, the lower the reference, the worse the outcome would have been had they made *another* choice ($O_{CF} < O_F$, Figure 12). Conversely, participants overestimating the true mean could be seen as pessimistic, as they assumed that the alternative course of action would have been better off on average ($O_{CF} > O_F$, Figure 12). This result agrees with a variety of empirical works showing that, while healthy adults exhibit attribution biases when judging their agency, these biases vary substantially across individuals (see Mezulis et al., 2004, for a review). By combining counterfactual updating with a subjective reference point, the CF model allows accounting for interindividual variability in self-serving beliefs, hence perceived controllability, during online instrumental learning.

Counterfactual emulation operates at most abstract levels of action control

Negative covariance is at the heart of the notion of "difference-making" in counterfactual theories of causal reasoning. Counterfactual (CF) theories posit that a cause is something that makes a difference to another event (Walsh and Sloman, 2011). According to the CF view, individuals would infer causal relations by simulating models of close alternatives ("nearest possible worlds") in which the candidate cause (A) is negated and the outcome is observed (O). If the outcome is undone (O) as a result of negating the candidate cause (A), then the probability that A is selected as the cause should increase accordingly (e.g., Roese, 1997; Sloman and Lagnado, 2015; Woodward, 2005). When applied to intentional causation, an action should be considered ?causal? if simulating a change in that action (e.g., the action is not taken) produces a change to the outcome (e.g., the outcome does not occur). In CF theories, "mutability" is a property that characterizes the effects of "simulating" changes in one variable, and hence can be seen as a measure of its causal power (Kahneman and Miller, 1986; Dehghani, Iliev, and Kaufmann, 2012). Thus, a variable is highly mutable if realizing, relative to not realizing, this variable is likely to make a difference to the world. Put differently: a variable is mutable if it is likely to diverge across factual and counterfactual worlds (Lucas and Kemp, 2015).

The notion of "mutability" closely relates to the notion of counterfactual learning, as instantiated in the CF model through a counterfactual (CF) learning rate parameter. A CF learning rate can be viewed as measuring the *speed of divergence* between factual and counterfactual outcomes. Thus, the greater the value of the CF learning rate, the faster the counterfactual action value is assumed to diverge from the chosen action value. In our task, a machine associated with a *high* CF learning rate is a *highly mutable* machine: choosing this machine, rather than the other one, is thought to make a significant difference with respect to the outcome. Importantly, our results revealed that counterfactual learning was hierarchically organized: CF learning rates were higher for the machine than for the button, suggesting that participants conceived the former as being more "mutable" than the latter. In other words, participants considered that making a choice about the machine was more likely to make a difference to the world relative to making a choice about the button.

This hierarchy in learning from factual and counterfactual action values might be well adapted to an environment where changes in causal influence (e.g., reversals) operate at more or less abstract levels of action control. Higher learning rates for the CF machine, relative to the CF button, suggest that subjects are more likely to engage in counterfactual emulation for testing their control at most abstract levels of action representation (the machine), and less for less abstract levels (the button) where *direct* instrumental testing is available. Should this prediction be correct, such hierarchy in counterfactual learning would be advantageous in an environment where instrumental divergence is maximal between machines, rather than between buttons. We directly tested this hypothesis by simulating our CF model across two different environments, where maximal instrumental divergence was either between buttons or between machines. In line with our predictions, we found that the CF model was best suited to deal with an environment where the divergence was set at the most abstract level, i.e., the machine level (see Figure 13A).

That individuals are more likely to engage in counterfactual emulation for the most abstract levels of action control is consistent with evidence from hierarchical models of action representation (e.g., Chambon et al., 2017; Kilner, 2011). In such models, an observer predicts another people's behaviour based on beliefs derived from simulating one's internal model (i.e., a model of how people are likely to behave in a given context). The nature of these beliefs critically depends on the level at which the behaviour is represented, from least to most abstract levels (e.g., kinematic vs. motor vs. goal level). Thus, a change at the most abstract level (e.g., the goal level: going to restaurant vs. theatre) is assumed to have a greater effect on the resulting behaviour than a change made at a less abstract level (e.g., the kinematic level: using a power vs. precision grip to grasp a mug). Importantly, human subjects show greater reliance on their internal models when having to predict more and more abstract behaviours (e.g., going to the restaurant vs. theatre > using a power vs. precision grip) (Chambon et al., 2017; Chambon et al., 2011). Likewise, our results indicate that human subjects are more likely to emulate counterfactual alternatives when making decision at more abstract levels of action control (machine > button).

Control beliefs foster opportunities for learning

Assuming a *negative* covariance between factual and counterfactual outcomes implies that the world can be divided into states that are essentially anti-correlated. In this scenario, two states only are possible: you are the agent, or you are not. This assumption agrees with the fact that judgments of agency are *binary* in nature. Indeed, while individuals readily experience intermediate levels of sensorimotor control, confidence, or difficulty, they rarely, if never, experience intermediate levels of agency; they can be "more or less confident", but they do not feel "more or less agent" (Chambon and Haggard, 2013). The all-or-none nature of agency is supported further by the observation that people think of causal relationships between actions and outcomes in terms of "state" (is A the cause of O?) rather than in terms of "force" (Tenenbaum and Griffiths, 2001). We argue that the CF model outperforms all other models in the set because it embeds a prior that matches the agentive structure of the current task, where binary and abrupt, rather than smooth and continuous, changes in contingencies could occur. In this sense, the CF model would be best suited to track changes in agency (me vs. not-me) than gradual changes due to external volatility (e.g., the light decreasing over the course of a day).

We speculate that this prior about negative covariance mirrors participants' belief about their control over machine outcomes: had their choice been different, the outcomes would also have been different. Importantly, participants hold this control belief even in sessions where no true control existed (see Experiment 1, "value" and "variance" sessions), or despite the fact that instructions made no reference to control in the task (see Experiment 1b, Supplementary information). Beliefs in one's causal power are a strong determinant of voluntary behaviours: individuals a more likely to enact certain behaviours when they feel or believe they can enact these behaviours successfully (e.g., Ajzen, 2002). Control beliefs develop early and are somewhat irrepressible: the need to be and feel in control is so strong that individuals would do whatever they can to re-establish control when it disappears or is taken away, including self-attributing unrelated events (Langer, 1975) or acting superstitiously in the belief that their action is accountable for uncontrollable outcomes (Blanco, Matute, and Vadillo, 2011). Importantly, control beliefs would explain an enduring puzzle in causal reasoning, that is, why people show remarkable performance in causal inferences, which they often make effortlessly and from very little data, and yet readily experience illusory control, whether in real-life uncontrollable situations (Matute, 1996) or in experimental settings with null contingency (Blanco, Matute, and Vadillo, 2011). This relationship between illusory control and control beliefs is further corroborated by people's tendency to self-attribute positive outcomes when perceived controllability of the environment is high (Harris and Osman, 2012).

One explanation for assuming control as a default belief – whether illusory or not – is *learning*. Indeed, control beliefs would be particularly adapted to controllable environments, whose latent causal structure can be learnt so as to maximise rewards in the long run

(Lake et al., 2017). In a structured environment, enacting actions, relative to not acting, is advantageous on average, as the reward/punishment ratio can be turned in favour of rewards though implementing appropriate actions. In this situation, agents would be better off holding the belief that their actions are efficient means for attaining desired outcomes. In sum, the detrimental consequences of assuming control as a default belief would be offset by opportunities for learning the causal structure of the world, and hence by the ability to flexibly switch preferences when reversals occur, ultimately reducing the cost associated with missing opportunities (Koechlin, 2016). Importantly, control beliefs, such as self-efficacy, play a major role in general health and wellbeing (Bobak et al., 2000). Lowered sense of causation is associated with lower self-esteem, greater mood disorders and greater depressive symptoms (Bandura, 1997). Depressed individuals perceive their environment as being more random than non-depressed people. In the depressed view (so-called "depressive realism"), the reward/punishment ratio is evenly balanced, which substantially reduces opportunities for learning and makes ultimately any action pointless (Nettle and Bateson, 2012). This account agrees with clinical reports of greater passivity in depressed people, that is, a reduced ability to initiate voluntary actions (Blanco, Matute, and Vadillo, 2012). Acting with less frequency would make depressed individuals exposed to fewer incidental associations between actions and action-contingent events (reduced "action-density" bias, see Matute et al., 2015), which might in turn impede learning of instrumental contingencies and aggravate depressive symptoms in the long run.

The strength of the CF model stems from the simplicity with which it embeds the participant's prior about control. This prior amounts to assuming relative (i.e., context-dependent) divergence between factual and counterfactual worlds. We argue that this simple prior allows the model to rapidly and flexibly switch preferences when a reversal occurs, as demonstrated by its robust learning curves and performances in both experiments, relative to more sophisticated models such as those aiming at statistical optimality (e.g., BC model). We believe that simplicity is required to account for the ease with which resource-bounded agents learn instrumental contingencies, but also to explain how strong control beliefs can be sustained as a default backdrop to our normal mental life. As mentioned above, a pervasive belief in one's causal power can make instrumental learning sometimes depart from statistical optimality, resulting in illusions of control and an inflated perception of one's own efficacy. The influence of such a pervasive belief would explain why learning of action-outcome causal relationships seems not to suffer the same biases as other forms of causal learning that are based on passive observation (Morris et al., 2017; Chambers et al., 2017). Ultimately, strong control beliefs in human agents could account for why reasoning about *external* causes differs from reflecting upon one's own causal power, both in terms of underlying computations, normative principles and optimal behaviour.

Conclusions

We designed a series of experiments that required participants to continuously monitor their causal influence over the task, through discriminating changes that were caused by their own actions from changes that were not. Comparing different models of choice, we found that participants' behaviour was best explained by a model (CF) deriving the consequences of the forgone action from the current action taken, and assuming relative divergence between both. Importantly, this model agrees with the intuitive way of construing causation as "difference-making", and further endorses the long-held view that goal-directed actions are divergent in nature: they make a difference to the world, and can hence be implemented as efficient means for pursuing desirable outcomes. In the CF model, difference-making was explicitly accounted for by assuming negative covariance between factual and counterfactual action values. Based on this covariance prior, the CF model directly emulated counterfactual action values through a subjective reference point that aligned with the actual structure of the task environment. Crucially, we found that counterfactual emulation was more likely to occur at most abstract levels of action control, consistent with evidence from hierarchical models of goal-directed actions. Finally, we suggest that the CF model outperformed all competitors because it closely mirrors people's belief in their causal power, a belief that is well suited to learning action-outcome associations in controllable environments. We speculate that control beliefs may be part of the reason why reflecting upon one's own causal power fundamentally differs from reasoning about external causes.

Acknowledgements

We thank Valentin Wyart, Stefano Palminteri, and Patrick Haggard for comments and useful discussions about earlier versions of this manuscript. V.C. was supported by the Agence Nationale pour la Recherche (ANR) grants ANR-10-LABX-0087 IEC (Institut d'Etude de la Cognition), ANR-10-IDEX-0001-02 PSL* (program "Investissements d'Avenir"), and ANR-16-CE37-0012-01. H.T. was supported by an PSL/ENS studentship. E.K. was supported by a European Research Council Grant (ERC-2009-AdG #250106). C.F. was supported by a DGA/INSERM fellowship.

Bibliography

Ajzen, Icek (2002). "Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior". In: *Journal of applied social psychology* 32.4, pp. 665–683.

Bandura, Albert (1989). "Human agency in social cognitive theory." In: *American psychologist* 44.9, p. 1175.

– (1997). *Self-efficacy: The exercise of control*. Macmillan.

Beebee, Helen, Christopher Hitchcock, and Huw Price (2017). *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press.

Bell, David E (1982). "Regret in decision making under uncertainty". In: *Operations research* 30.5, pp. 961–981.

Berlyne, Daniel E (1950). "Novelty and curiosity as determinants of exploratory behaviour". In: *British Journal of Psychology* 41.1-2, pp. 68–80.

– (1966). "Conflict and arousal". In: *Scientific American* 215.2, pp. 82–87.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

Bjork, James M and Daniel W Hommer (2007). "Anticipating instrumentally obtained and passively-received rewards: a factorial fMRI investigation". In: *Behavioural brain research* 177.1, pp. 165–170.

Blanco, Fernando, Helena Matute, and Miguel A Vadillo (2011). "Making the uncontrollable seem controllable: The role of action in the illusion of control". In: *Quarterly Journal of Experimental Psychology* 64.7, pp. 1290–1304.

– (2012). "Mediating role of activity level in the depressive realism effect". In: *PLoS One* 7.9, e46203.

Bobak, Martin et al. (2000). "Socioeconomic factors, material inequalities, and perceived control in self-rated health: cross-sectional data from seven post-communist countries". In: *Social science & medicine* 51.9, pp. 1343–1350.

Boland, Philip J and Yudi Pawitan (1999). "Trying to be random in selecting numbers for Lotto". In: *Journal of Statistics Education* 7.3.

Boorman, Erie D, Timothy E Behrens, and Matthew F Rushworth (2011). "Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex". In: *PLoS biology* 9.6, e1001093.

Bowers, Jeffrey S and Colin J Davis (2012). "Bayesian just-so stories in psychology and neuroscience." In: *Psychological bulletin* 138.3, p. 389.

Bown, Nicola J, Daniel Read, and Barbara Summers (2003). "The lure of choice". In: *Journal of Behavioral Decision Making* 16.4, pp. 297–308.

Bramley, Neil R et al. (2017). "Formalizing Neurath's ship: Approximate algorithms for online causal learning." In: *Psychological review* 124.3, p. 301.

Brehm, Jack W (1966). *A theory of psychological reactance.* Academic Press.

Brehm, Sharon S and Jack W Brehm (1981). *Psychological reactance: A theory of freedom and control*. Academic Press.

Buchsbaum, Daphna et al. (2012). "The power of possibility: Causal learning, counterfactual reasoning, and pretend play". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1599, pp. 2202–2212.

Burger, J. M. (2016). "And the Wisdom to Know the Difference: Locus of Control and Desire for Control." In: *J. W. Reich & F. J. Infurna (Eds.). Perceived control: Theory, Research and Practice in the first 50 years. Oxford, England: Oxford University Press*, 45–70.

Burke, Christopher J et al. (2016). "Partial adaptation of obtained and observed value signals preserves information about gains and losses". In: *Journal of Neuroscience* 36.39, pp. 10016–10025.

Butler, Robert A (1953). "Discrimination learning by rhesus monkeys to visual-exploration motivation." In: *Journal of Comparative and Physiological Psychology* 46.2, p. 95.

Byrne, Ruth MJ (2002). "Mental models and counterfactual thoughts about what might have been". In: *Trends in cognitive sciences* 6.10, pp. 426–431.

– (2005). *The rational imagination: How people create alternatives to reality*. MIT press.

Caruso, Eugene M, Adam Waytz, and Nicholas Epley (2010). "The intentional mind and the hot hand: Perceiving intentions makes streaks seem likely to continue". In: *Cognition* 116.1, pp. 149–153.

Chambers, Claire et al. (2017). "The development of Bayesian integration in sensorimotor estimation". In: *bioRxiv*, p. 136267.

Chambon, Valerian and Patrick Haggard (2012). "Sense of control depends on fluency of action selection, not motor performance". In: *Cognition* 125.3, pp. 441–451.

– (2013). "Premotor or Ideomotor: How Does the Experience of Action Come About?" In: *Action science: Foundations of an emerging discipline*, p. 359.

Chambon, Valerian et al. (2011). "What are they up to? The role of sensory evidence and prior knowledge in action understanding". In: *PloS one* 6.2, e17133.

Chambon, Valerian et al. (2017). "Neural coding of prior expectations in hierarchical intention inference". In: *Scientific Reports* 7.1, p. 1278.

Cheng, Patricia W (1997). "From covariation to causation: a causal power theory." In: *Psychological review* 104.2, p. 367.

Cockburn, Jeffrey, Anne GE Collins, and Michael J Frank (2014). "A reinforcement learning mechanism responsible for the valuation of free choice". In: *Neuron* 83.3, pp. 551–557.

Coricelli, Giorgio et al. (2005). "Regret and its avoidance: a neuroimaging study of choice behavior". In: *Nature neuroscience* 8.9, p. 1255.

Daunizeau, Jean, Vincent Adam, and Lionel Rigoux (2014). "VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data". In: *PLoS computational biology* 10.1, e1003441.

Daw, Nathaniel D and Peter Dayan (2014). "The algorithmic anatomy of model-based evaluation". In: *Phil. Trans. R. Soc. B* 369.1655, p. 20130478.

Daw, Nathaniel D, Yael Niv, and Peter Dayan (2005). "Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control". In: *Nature neuroscience* 8.12, p. 1704.

Daw, Nathaniel D et al. (2006). "Cortical substrates for exploratory decisions in humans". In: *Nature* 441.7095, p. 876.

Dayan, Peter (1991). "Reinforcement comparison". In: *Connectionist Models*. Elsevier, pp. 45–51.

De Gardelle, Vincent and Christopher Summerfield (2011). "Robust averaging during perceptual judgment". In: *Proceedings of the National Academy of Sciences* 108.32, pp. 13341–13346.

Dehghani, Morteza, Rumen Iliev, and Stefan Kaufmann (2012). "Causal explanation and fact mutability in counterfactual reasoning". In: *Mind & Language* 27.1, pp. 55–85.

Denrell, Jerker C (2015). "Reference-dependent risk sensitivity as rational inference." In: *Psychological review* 122.3, p. 461.

Dezfouli, Amir and Bernard W Balleine (2013). "Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized". In: *PLoS computational biology* 9.12, e1003364.

Dickinson, Anthony (2001). "The 28th Bartlett memorial lecture causal learning: an associative analysis". In: *The Quarterly Journal of Experimental Psychology Section B* 54.1b, pp. 3–25.

Doya, Kenji et al. (2002). "Multiple model-based reinforcement learning". In: *Neural computation* 14.6, pp. 1347–1369.

Eckstein, Miguel P et al. (2004). "Perceptual learning through optimization of attentional weighting: Human versus optimal Bayesian learner". In: *Journal of Vision* 4.12, pp. 3–3.

Elsner, Birgit and Bernhard Hommel (2001). "Effect anticipation and action control." In: *Journal of experimental psychology: human perception and performance* 27.1, p. 229.

Gershman, Samuel J (2015). "A unifying probabilistic view of associative learning". In: *PLoS computational biology* 11.11, e1004567.

Gershman, Samuel J, Arthur B Markman, and A Ross Otto (2014). "Retrospective revaluation in sequential decision making: A tale of two systems." In: *Journal of Experimental Psychology: General* 143.1, p. 182.

Girotto, Vittorio, Paolo Legrenzi, and Antonio Rizzo (1991). "Event controllability in counterfactual thinking". In: *Acta Psychologica* 78.1-3, pp. 111–133.

Gläscher, Jan et al. (2010). "States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning". In: *Neuron* 66.4, pp. 585–595.

Gopnik, Alison, Laura Schulz, and Laura Elizabeth Schulz (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Guruprasad, Shashi, Robert Ricci, and Jay Lepreau (2005). "Integrated network experimentation using simulation and emulation". In: *Testbeds and Research Infrastructures for the Development of Networks and Communities, 2005. Tridentcom 2005. First International Conference on*. IEEE, pp. 204–212.

Haggard, Patrick and Valerian Chambon (2012). "Sense of agency". In: *Current Biology* 22.10, R390–R392.

Harris, Adam JL and Magda Osman (2012). "The illusion of control: A Bayesian perspective". In: *Synthese* 189.1, pp. 29–38.

Harris, Peter (1996). "Sufficient grounds for optimism?: The relationship between perceived controllability and optimistic bias". In: *Journal of Social and Clinical Psychology* 15.1, pp. 9–52.

Heckhausen, Jutta and Richard Schulz (1995). "A life-span theory of control." In: *Psychological review* 102.2, p. 284.

Heider, Fritz (1958). *The psychology of interpersonal relations*. New York: Wiley.

Helwig, Charles C (2006). "The development of personal autonomy throughout cultures". In: *Cognitive Development* 21.4, pp. 458–473.

Hendrick, Ives (1943). "The Discussion of the 'Instinct to Master: A Letter to the Editors". In: *The Psychoanalytic Quarterly* 12.4, pp. 561–565.

Hume, D. (1748). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals.* Vol. 1. 1. Oxford, England: Clarendon Press.

Illari, Phyllis McKay, Federica Russo, and Jon Williamson (2011). *Causality in the Sciences*. Oxford University Press.

Jones, Matt and Bradley C Love (2011). "Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition". In: *Behavioral and Brain Sciences* 34.4, pp. 169–188.

Kaelbling, Leslie Pack, Michael L Littman, and Andrew W Moore (1996). "Reinforcement learning: A survey". In: *Journal of artificial intelligence research* 4, pp. 237–285.

Kahneman, Daniel and Dale T Miller (1986). "Norm theory: Comparing reality to its alternatives." In: *Psychological review* 93.2, p. 136.

Karsh, N and B Eitam (2015). "I control therefore I do: Judgments of agency influence action selection". In: *Cognition* 138, pp. 122–131.

Keramati, Mehdi, Amir Dezfouli, and Payam Piray (2011). "Speed/accuracy trade-off between the habitual and the goal-directed processes". In: *PLoS computational biology* 7.5, e1002055.

Kilner, James M (2011). "More than one pathway to action understanding". In: *Trends in cognitive sciences* 15.8, pp. 352–357.

Klein, Tilmann A, Markus Ullsperger, and Gerhard Jocham (2017). "Learning relative values in the striatum induces violations of normative decision making". In: *Nature communications* 8, p. 16033.

Koechlin, Etienne (2014). "An evolutionary computational theory of prefrontal executive function in decision-making". In: *Phil. Trans. R. Soc. B* 369.1655, p. 20130474.

– (2016). "Prefrontal executive function and adaptive behavior in complex environments". In: *Current opinion in neurobiology* 37, pp. 1–6.

Kool, Wouter, Sarah J Getz, and Matthew M Botvinick (2013). "Neural representation of reward probability: evidence from the illusion of control". In: *Journal of cognitive neuroscience* 25.6, pp. 852–861.

Kőszegi, Botond and Matthew Rabin (2006). "A model of reference-dependent preferences". In: *The Quarterly Journal of Economics* 121.4, pp. 1133–1165.

Lake, Brenden M et al. (2017). "Building machines that learn and think like people". In: *Behavioral and Brain Sciences* 40.

Langer, Ellen J (1975). "The illusion of control." In: *Journal of personality and social psychology* 32.2, p. 311.

Lau, Brian and Paul W Glimcher (2005). "Dynamic response-by-response models of matching behavior in rhesus monkeys". In: *Journal of the experimental analysis of behavior* 84.3, pp. 555–579.

Leotti, Lauren A, Sheena S Iyengar, and Kevin N Ochsner (2010). "Born to choose: The origins and value of the need for control". In: *Trends in cognitive sciences* 14.10, pp. 457–463.

Liljeholm, Mimi et al. (2011). "Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction". In: *Journal of Neuroscience* 31.7, pp. 2474–2480.

Liljeholm, Mimi et al. (2013). "Neural correlates of the divergence of instrumental probability distributions". In: *Journal of Neuroscience* 33.30, pp. 12519–12527.

Lohrenz, Terry et al. (2007). "Neural signature of fictive learning signals in a sequential investment task". In: *Proceedings of the National Academy of Sciences* 104.22, pp. 9493–9498.

Lucas, Christopher G and Charles Kemp (2015). "An improved probabilistic account of counterfactual reasoning." In: *Psychological review* 122.4, p. 700.

Maier, Steven F and Martin EP Seligman (2016). "Learned helplessness at fifty: Insights from neuroscience." In: *Psychological review* 123.4, p. 349.

Markman, Arthur B and A Ross Otto (2011). "Cognitive systems optimize energy rather than information". In: *Behavioral and Brain Sciences* 34.4, pp. 207–207.

Matute, Helena (1996). "Illusion of control: Detecting response-outcome independence in analytic but not in naturalistic conditions". In: *Psychological Science* 7.5, pp. 289–293.

Matute, Helena et al. (2015). "Illusions of causality: how they bias our everyday thinking and how they could be reduced". In: *Frontiers in Psychology* 6, p. 888.

McCloy, Rachel and Ruth MJ Byrne (2002). "Semifactual "even if" thinking". In: *Thinking & Reasoning* 8.1, pp. 41–67.

Mezulis, Amy H et al. (2004). "Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias." In: *Psychological bulletin* 130.5, p. 711.

Mineka, Susan and Robert W Hendersen (1985). "Controllability and predictability in acquired motivation". In: *Annual review of psychology* 36.1, pp. 495–529.

Mistry, Prachi and Mimi Liljeholm (2016). "Instrumental Divergence and the Value of Control". In: *Scientific reports* 6, p. 36295.

Morris, Richard W et al. (2017). "The algorithmic neuroanatomy of action-outcome learning". In: *bioRxiv*, p. 137851.

Nettle, Daniel and Melissa Bateson (2012). "The evolutionary origins of mood and its disorders". In: *Current Biology* 22.17, R712–R721.

Nissen, Henry W (1930). "A study of exploratory behavior in the white rat by means of the obstruction method". In: *The Pedagogical Seminary and Journal of Genetic Psychology* 37.3, pp. 361–376.

N'Gbala, Ahogni and Nyla R Branscombe (1995). "Mental simulation and causal attribution: When simulating an event does not affect fault assignment". In: *Journal of Experimental Social Psychology* 31.2, pp. 139–162.

O'Doherty, John et al. (2004). "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". In: *science* 304.5669, pp. 452–454.

O'Reilly, Jill X et al. (2013). "Brain systems for probabilistic and dynamic prediction: computational specificity and integration". In: *PLoS biology* 11.9, e1001662.

Palminteri, Stefano et al. (2015). "Contextual modulation of value signals in reward and punishment learning". In: *Nature communications* 6, p. 8096.

Palminteri, Stefano et al. (2017). "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8, e1005684.

Parpart, Paula et al. (2017). "Active learning reveals underlying decision strategies". In: *bioRxiv*, p. 239558.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. Cambridge university press.

Pittman, Thane S and Nancy L Pittman (1980). "Deprivation of control and the attribution process." In: *Journal of Personality and Social Psychology* 39.3, p. 377.

Premack, David (2007). "Human and animal cognition: Continuity and discontinuity". In: *Proceedings of the National Academy of Sciences* 104.35, pp. 13861–13867.

Redgrave, Peter and Kevin Gurney (2006). "The short-latency dopamine signal: a role in discovering novel actions?" In: *Nature reviews neuroscience* 7.12, p. 967.

Rescorla, Robert A, Allan R Wagner, et al. (1972). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement". In: *Classical conditioning II: Current research and theory* 2, pp. 64–99.

Roese, Neal J (1997). "Counterfactual thinking." In: *Psychological bulletin* 121.1, p. 133.

Rolls, Edmund T (2000). "The orbitofrontal cortex and reward". In: *Cerebral cortex* 10.3, pp. 284–294.

Rothbaum, Fred, John R Weisz, and Samuel S Snyder (1982). "Changing the world and changing the self: A two-process model of perceived control." In: *Journal of personality and social psychology* 42.1, p. 5.

Rubenstein, LIZA M, LAUREN B Alloy, and LYNY Abramson (2016). "Perceived Control and Depression". In: *Perceived Control: Theory, Research, and Practice in the First 50 Years*, p. 229.

Ryan, Richard M and Edward L Deci (2006). "Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will?" In: *Journal of personality* 74.6, pp. 1557–1586.

Schneider, Kurt (1959). "Clinical psychopathology.(Trans. by MW Hamilton)". In:

Scholl, Annika and Kai Sassenberg (2014). "Where could we stand if I had. . . ? How social power impacts counterfactual thinking after failure". In: *Journal of Experimental Social Psychology* 53, pp. 51–61.

Seymour, Ben and Samuel M McClure (2008). "Anchors, scales and the relative coding of value in the brain". In: *Current opinion in neurobiology* 18.2, pp. 173–178.

Shapiro, Jr Deane H, Carolyn E Schwartz, and John A Astin (1996). "Controlling ourselves, controlling our world: Psychology's role in understanding positive and negative consequences of seeking and gaining control." In: *American psychologist* 51.12, p. 1213.

Sharot, Tali, Benedetto De Martino, and Raymond J Dolan (2009). "How choice reveals and shapes expected hedonic outcome". In: *Journal of Neuroscience* 29.12, pp. 3760–3765.

Sharot, Tali, Tamara Shiner, and Raymond J Dolan (2010). "Experience and choice shape expected aversive outcomes". In: *Journal of Neuroscience* 30.27, pp. 9209–9215.

Siwak, Christina T, P Dwight Tapp, and Norton W Milgram (2001). "Effect of age and level of cognitive function on spontaneous and exploratory behaviors in the beagle dog". In: *Learning & memory* 8.6, pp. 317–325.

Sloman, Steven A and David Lagnado (2015). "Causality in thought". In: *Annual Review of Psychology* 66, pp. 223–247.

Solway, Alec and Matthew M Botvinick (2015). "Evidence integration in model-based tree search". In: *Proceedings of the National Academy of Sciences* 112.37, pp. 11708–11713.

Spellman, Barbara A and Alexandra Kincannon (2001). "The relation between counterfactual (" but for") and causal reasoning: Experimental findings and implications for jurors' decisions". In: *Law and Contemporary Problems* 64.4, pp. 241–264.

Stefan, Simona and Daniel David (2013). "Recent developments in the experimental investigation of the illusion of control. A meta-analytic review". In: *Journal of Applied Social Psychology* 43.2, pp. 377–386.

Sutton, Richard S and Andrew G Barto (1998). *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge.

Sutton, Richard Stuart (1984). *Temporal credit assignment in reinforcement learning*. University of Massachussets, Amherst, MA.

Suzuki, Shuji (1999). "Selection of forced-and free-choice by monkeys (Macaca fascicularis)". In: *Perceptual and motor skills* 88.1, pp. 242–250.

Tanaka, Saori C, Bernard W Balleine, and John P O'Doherty (2008). "Calculating consequences: brain systems that encode the causal effects of actions". In: *Journal of Neuroscience* 28.26, pp. 6750–6755.

Tenenbaum, Joshua B and Thomas L Griffiths (2001). "Structure learning in human causal induction". In: *Advances in neural information processing systems*, pp. 59–65.

Tricomi, Elizabeth M, Mauricio R Delgado, and Julie A Fiez (2004). "Modulation of caudate activity by action contingency". In: *Neuron* 41.2, pp. 281–292.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under uncertainty: Heuristics and biases". In: *science* 185.4157, pp. 1124–1131.

Waldmann, Michael R and York Hagmayer (2005). "Seeing versus doing: two modes of accessing causal knowledge." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.2, p. 216.

Waldmann, Michael R and Jessica M Walker (2005). "Competence and performance in causal learning". In: *Learning & Behavior* 33.2, pp. 211–229.

Walsh, Clare R and Steven A Sloman (2011). "The meaning of cause and prevent: The role of causal mechanism". In: *Mind & Language* 26.1, pp. 21–52.

Ward, William C and Herbert M Jenkins (1965). "The display of information and the judgment of contingency." In: *Canadian Journal of Psychology/Revue canadienne de psychologie* 19.3, p. 231.

White, Robert W (1959). "Motivation reconsidered: The concept of competence." In: *Psychological review* 66.5, p. 297.

Whitson, Jennifer A and Adam D Galinsky (2008). "Lacking control increases illusory pattern perception". In: *science* 322.5898, pp. 115–117.

Woodward, James (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Zimbardo, Philip G and Neal E Miller (1958). "Facilitation of exploration by hunger in rats." In: *Journal of Comparative and Physiological Psychology* 51.1, p. 43.

Supplementary material

Appendix A. Generative model

As mentioned in the main text, the generative model was a Bayesian learner that updated beliefs associated with each possible action, on each trial. Specifically, the model (either BM or BC) aims to infer the correct action mapping between the four possible mappings (or states). We further define a state-function f specifying the underlying reward distribution for a given action a and state z, as follows:

State 1: $f(a=1,z=1)=G_1; f(a=2,z=1)=G_2; f(a=3,z=1)=G_3; f(a=4,z=1)=G_3.$
State 2: $f(a=1,z=2)=G_2; f(a=2,z=2)=G_1; f(a=3,z=2)=G_3; f(a=4,z=2)=G_3.$
State 3: $f(a=1,z=3)=G_3; f(a=2,z=3)=G_3; f(a=3,z=3)=G_1; f(a=4,z=3)=G_2.$
State 4: $f(a=1,z=4)=G_3; f(a=2,z=4)=G_3; f(a=3,z=4)=G_2; f(a=4,z=4)=G_1.$

where a corresponds to each possible action (among the 4 possible combinations of button and machine), G_1 is the distribution associated with having selected the best-rewarding button of the controlled machine, G_2 is the distribution associated with having selected the least-rewarding button of the controlled machine, and G_3 is the distribution associated with having selected the non-controlled machine (see respectively green, red, and grey distributions of Figure 4, top panel). We assume the rewards to be drawn from Gaussian distributions, as they were in the task (see Figure 5A).

The analytical model used to infer the state on each trial is a hidden Markov model, defined as follows:

$$\begin{aligned} z_1 &\sim Unif(\{1,2,3,4\}) \\ z_t|z_{t-1},\tau &\sim (1-\tau)\times\delta_{z_t z_{t-1}} + \tau\times Unif(k\in\{1,2,3,4\},k\neq z_{t-1}) \\ r_t|a_t,z_t,\mu_1,\mu_2,\mu_3,\lambda_1,\lambda_2,\lambda_3 &\sim Norm(r_t|\mu_{f(a_t,z_t)},\lambda_{f(a_t,z_t)}) \end{aligned}$$

where z_t corresponds to the state inferred on trial t, r_t corresponds to the reward observed on trial t, a_t corresponds to the action observed on trial t, and δ is the index function, i.e., $\delta_{ab} = \begin{cases} 1 \text{ if } a=b \\ 0 \text{ if } a\neq b \end{cases}$.

Note that the reward values were rescaled within the range [0,1].

Let the parameters be $\theta=(\tau,\mu_1,\mu_2,\mu_3,\lambda_1,\lambda_2,\lambda_3)$. The analytical model to infer the parameters on the first trial, given their prior hyperparameters, is the following:

$$\begin{aligned} \tau &\sim Beta(a^0,b^0) \\ p(\mu_1|\mu_2) &\propto Norm(\mu_1|\mu_1^0,\lambda_1^0)\times 1[\mu_1>\mu_2] \\ p(\mu_2|\mu_1) &\propto Norm(\mu_2|\mu_2^0,\lambda_2^0)\times 1[\mu_1>\mu_2] \\ \mu_3 &\sim Norm(\mu_3^0,\lambda_3^0) \\ \lambda_g &\sim Gamma(\alpha_0,\beta_0),\ \text{for } g\in\{1,2,3\} \end{aligned}$$

where τ is the volatility parameter, and μ_g and λ_g are the mean and the precision of the Gaussian G_g for $g\in\{1,2,3\}$. To obtain conjugate distributions, we used Gaussian precisions rather than their standard deviations. We assumed the same hyper-parameters for the precisions of the three Gaussians. These hyper-parameters led to

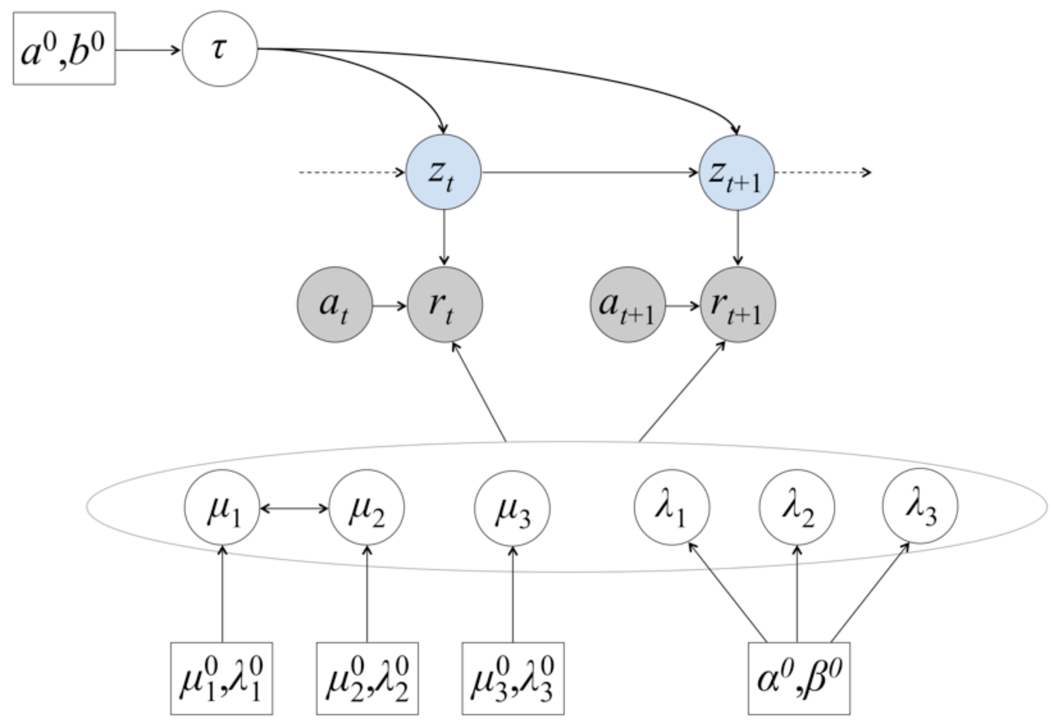


Figure S1: ***Generative graphical model assumed by the subject.*** *The variable z_t corresponds to the state on trial t (shown in light blue). r_t and a_t are the observed variables (shown in grey): r_t corresponds to the reward observed on trial t, and a_t corresponds to the action observed on trial t. The parameters are $\theta = (\tau, \mu_1, \mu_2, \mu_3, \lambda_1, \lambda_2, \lambda_3)$, shown in white circles: τ is the volatility parameter, and μ_g and λ_g are the mean and the precision of the Gaussian G_g for $g \in \{1, 2, 3\}$. The hyperparameters are shown in white boxes.*

the less informative priors possible and were set to the following values: $a^0 = 1, b^0 = 9, \mu_1^0 = 0.58, \mu_2^0 = 0.42, \mu_3^0 = 0.50, \lambda_1^0 = \lambda_2^0 = \lambda_3^0 = 20, \alpha_0 = 80, \beta_0 = 0.8$.

The goal of the inference algorithm is to predict the state of the next trial:

$$p(z_{t+1}|a_{1:t}, r_{1:t})$$

Let us have $I = 1,000$ samples $(z_{t+1}^i)_{i\in\{1,...,I\}}$ distributed under the distribution $p(z_{t+1}|a_{1:t}, r_{1:t})$. A Monte Carlo approximation of the integral gives:

$$p(z_{t+1} = k|a_{1:t}, r_{1:t}) \propto \frac{1}{I}\sum_{i=1}^{I} \delta_{kz_{t+1}^i}$$

To perform inference, we used a Gibbs algorithm sampling iteratively the latent states and the parameters.

Regarding the sampling of the latent states, for the first trial, the posterior distribution on the hidden state takes the following form:

$$p(z_1|\theta, a_1, r_1) \propto Norm(r_1|\mu_{f(a_1,z_1)}, \lambda_{f(a_1,z_1)})p(z_1)$$

Then, the forward recursion from trial $i-1$ to trial i leads to:

$$p(z_{i-1}, z_i|\theta, a_{1:i}, r_{1:i}) \propto Norm(r_i|\mu_{f(a_i,z_i)}, \lambda_{f(a_i,z_i)}) \times p(z_i|z_{i-1}, \tau) \times p(z_{i-1}|\theta, a_{1:i-1}, r_{1:i-1})$$

with $p(z_{i-1}|\theta, a_{1:i-1}, r_{1:i-1}) = \sum\limits_{z_{i-2}=1}^{4} p(z_{i-2}, z_{i-1}|\theta, a_{1:i-1}, r_{1:i-1})$.

The latent sample $z_{1:t}$ is thus obtained by drawing z_t from $p(z_t|\theta, a_{1:t}, r_{1:t})$, and then by iteratively sampling backward $z_{i-1}|z_i \sim p(., z_i|\theta, a_{1:i}, r_{1:i})$ (Scott, 2002).

For the parameter sampling step, the posterior distribution of the volatility parameter τ depends on the number of switches predicted by the sampling trajectory $z_{1:t}$. Let us define $N_{switch} = Card(i \in \{2, ..., t\} \mid z_{i-1} \neq z_i)$. The posterior distribution of the volatility parameter is updated as follows:

$$\tau|z_{1:t}, a_{1:t}, r_{1:t} \sim Beta(a^0 + N_{switch}, b^0 + t - 1 - N_{switch})$$

The means and precisions of the three Gaussians are also updated based on past history of actions and rewards. This update first requires identifying which Gaussian g each observed reward was drawn from. Let us define $I_g = \{i \in \{1, ..., t\} \mid f(a_i, z_i) = G_g\}$. Then, the number of trials in which the rewards are drawn from the Gaussian g is: $N_g = Card(I_g)$, and the average reward observed for the Gaussian g is:

$$\overline{r}_g = \frac{1}{N_g} \sum_{i \in I_g} r_i$$

To sample from the mean of G_3, the Gaussian associated with the non-controlled machine, one just computes the tractable posterior and samples from it:

$$\mu_3 | z_{1:t}, a_{1:t}, r_{1:t} \sim Norm(\frac{N_3 \lambda_3 \overline{r}_3 + \lambda_3^0 \mu_3^0}{N_3 \lambda_3 + \lambda_3^0},\ N_3 \lambda_3 + \lambda_3^0)$$

For the means of G_1 and G_2, the Gaussians associated with the best- and the least-rewarding button of the controlled machine, the additional inequality constraint makes the posterior distribution intractable. We thus us Monte Carlo Markov Chain procedures within the Gibbs algorithm to sample from the constrained conditional distributions of μ_1 and μ_2. For the proposal distribution of the Metropolis-Hastings algorithms implemented, we use the unconstrained posterior:

$$\mu_g | z_{1:t}, a_{1:t}, r_{1:t} \sim Norm(\frac{N_g \lambda_g \overline{r}_g + \lambda_g^0 \mu_g^0}{N_g \lambda_g + \lambda_g^0},\ N_g \lambda_g + \lambda_g^0),\ \text{for } g \in \{1, 2\}$$

As for the posterior distributions of the precision parameters λ_g, they are of the form:

$$\lambda_g | z_{1:t}, a_{1:t}, r_{1:t} \sim Gamma(\alpha_0 + \frac{N_g}{2},\ \beta_0 + \frac{1}{2} \sum_{i \in I_g} (r_i - \mu_g)^2),\ \text{for } g \in \{1, 2, 3\}$$

Appendix B. Experiment 1b: method and results

20 participants (11 females, age 21-29) took part in the experiment. The task (stimuli, timeline, and trial structure) was identical to that used in experiment 1. The only differences were the instructions and the order of sessions. The verbal and written instructions did not make any reference to a controlled machine or to a best-rewarding button. Thus, participants were only instructed to choose the option that would maximize their total reward while being reminded that they would always win the sum of the two machines on each trial. The order of the three experimental sessions was fully randomized so that participants could begin by any of the three sessions (dependency, value, or variability).

Experiment 1b comprised the same number of episodes as in experiment 1, with similar length and same number of button and/or machine reversals. The same four models (RL, CF, BC, BM) were fitted to participants' choices and simulated over all three sessions with their best-fitting parameters.

The results replicated the experiment 1. Participants discriminated well between the two buttons of the controlled machine in the dependency session only (best- and least-rewarding buttons: 0.33 vs. 0.20, $t_{19} = 3.96, p < 0.001$) while choosing equally button 1 and button 2 of the non-controlled machines in all sessions (all $t_{19} < 2.16$, all $p > 0.12$). Second, we tested whether participants showed a preference for one machine over another within each session, by comparing choice proportion for each machine against the chance level (0.5). As again expected, we found that participants showed a significant preference for the *controlled* machine in the dependency session ($t_{19} = 1.98, p < 0.05$) and a marked preference for the *best-rewarding* machine in the value condition ($t_{19} = 5.97, p < 0.001$). In contrast to experiment 1, however, they only showed a tendency to prefer the *low-variable* machine in the variance condition ($t_{19} = 1.2, p = 0.19$).

As in the previous experiment, we compared "button" preferences across all sessions by subtracting choice for one button from choice for the other button within each preferred machine. The one-way ANOVA confirmed that button preferences differed across the 3 sessions ($F(2, 57) = 12.23, p < 0.001, \eta_p^2 = 0.30$). Thus, participants discriminated between buttons of the preferred machine in the dependency session to a greater extent than in the value and variance sessions (post hoc tests: all $p < 0.001$). We then compared the proportion of choice for the preferred machine across all 3 sessions. The one-way ANOVA revealed that "machine" preferences differed across the 3 sessions ($F(2, 57) = 17.95, p < 0.001, \eta_p^2 = 0.38$), Thus, participants chose the best-rewarding machine (value session, 0.65) more than the controlled machine (dependency session, 0.54), and more than the low-variable machine (variance session, 0.52) (post hoc tests, all $p < 0.001$). Finally, participants were able to adjust to machine and/or button reversals, on average reaching the plateau of performance around 10 trials after reversal.

Again, replicating experiment 1, the CF model best predicted participants' choices than all other models in the set (RL, BM or BC), in all three sessions (exceedance probability > 92%) (see Table 1).

Most of the results from the previous experiment were replicated. As expected, participants exhibited a strong preference for high rewards and preferentially chose low-variable machines, regardless of their overall value. We also found that participants were able to discriminate between causally efficient actions (the two buttons) and to identify where in the task environment choosing one action rather than another made a difference to the outcome (the controlled machine), and where it did not (the non-controlled machine). Importantly, this pattern of choice preference could neither be explained by the session order nor by the instructions. In this experiment, participants could as well start with the dependency as with the value or the variability sessions. Furthermore, participants were only instructed to make choices that maximized their rewards over both machines and time. Finally, in this experiment as in the previous one, we found that a model based on a simple context-dependent counterfactual rule (CF) outperformed all competing models, including a pure reinforcement learner (RL) and a model that explicitly aimed at maximizing reward by means of Bayesian inference (BM).

Appendix C. Experiment 2: supplementary figures

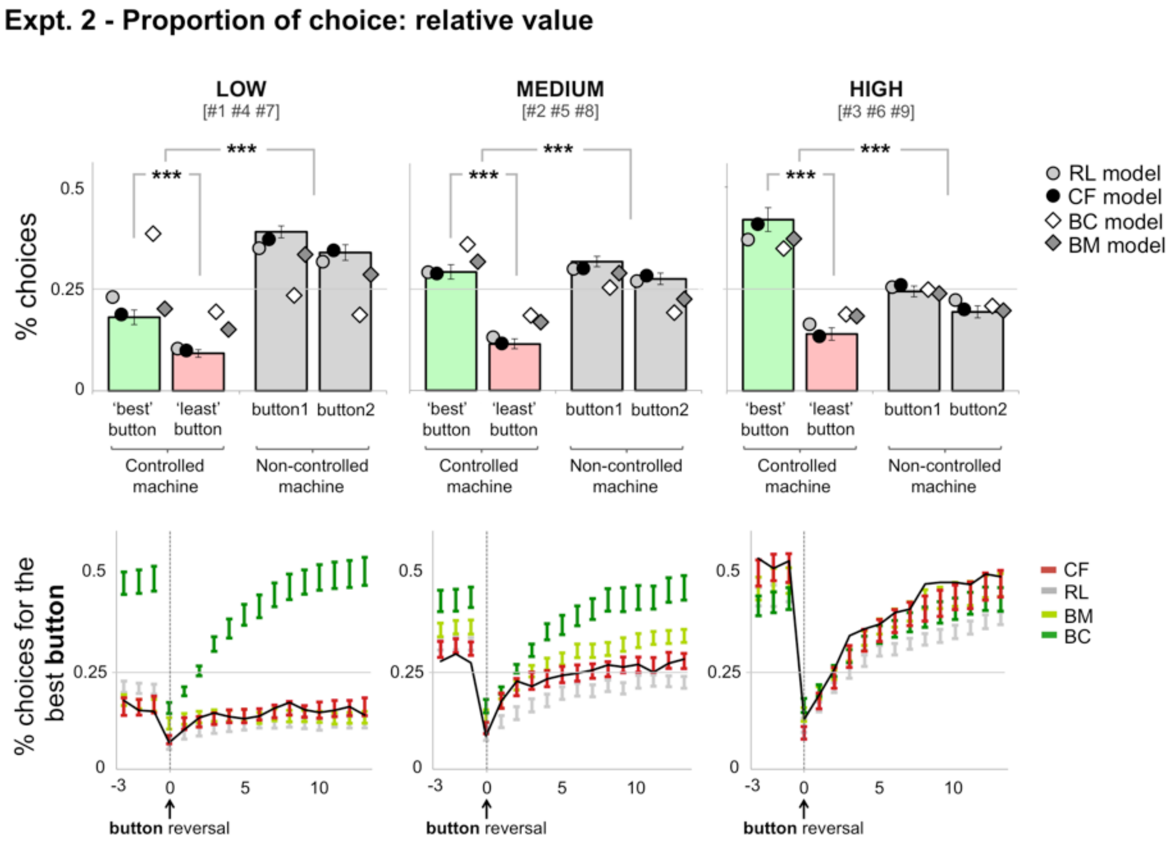


Figure S2: ***Top panel: Mean proportion of choice for each button of each button across the "value" dimension.*** *Low, medium, high: the value of the controlled machine was low, medium, or high, relative to the value of the non-controlled machine. The numbers between brackets refer to experimental conditions shown in Figure 5. All error bars indicate standard error. For the sake of visibility, models' error bars are not shown. Three-stars:* $p < 0.001$. ***Bottom panel: Reversal curves for participants (solid black line) and models (colored bars), across all three levels of the "value" dimension.*** *Model simulation: CF (red bars); RL (light grey); BM (light green bars); BC (dark green bars). Bars indicate standard error. Dashed vertical lines indicate reversal point. Machine reversal curves are not shown.*

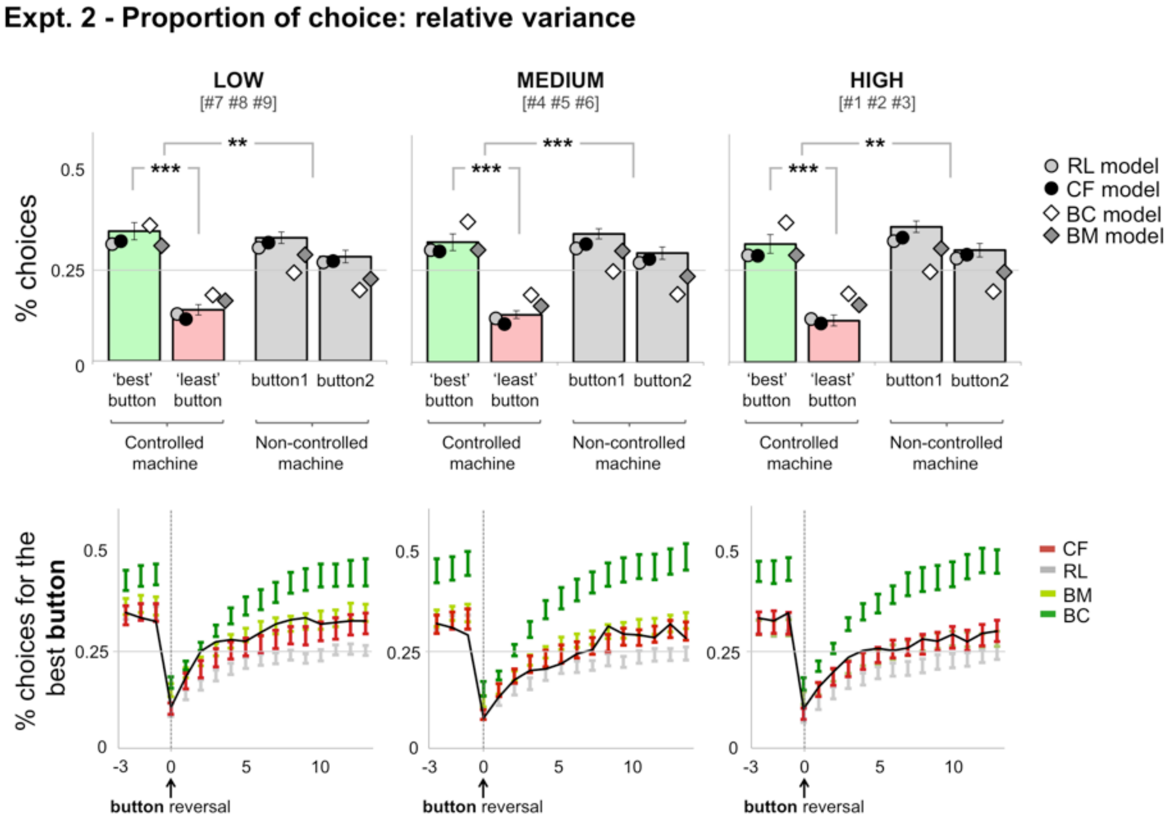


Figure S3: ***Top panel: Mean proportion of choice for each button of each button across the "variance" dimension.*** *Low, medium, high: the variance of the controlled machine was low, medium, or high, relative to the variance of the non-controlled machine. The numbers between brackets refer to experimental conditions shown in Figure 5. All error bars indicate standard error. For the sake of visibility, models' error bars are not shown. Two-stars: $p < 0.01$; Three-stars: $p < 0.001$.* ***Bottom panel: Reversal curves for participants (solid black line) and models (colored bars), across all three levels of the "variance" dimension.*** *Model simulation: CF (red bars); RL (light grey); BM (light green bars); BC (dark green bars). Bars indicate standard error. Dashed vertical lines indicate reversal point. Machine reversal curves are not shown.*

Appendix D. CF model simulations

Parameter recovery procedure

We used simulations to verify that the pattern of learning rates obtained in Experiments 1 and 2 did not arise artificially from the parameter optimization procedure. We ran a parameter recovery analysis for discrete sets of parameter values. For Experiments 1 and 2, we simulated 36 virtual participants on our behavioural tasks (36 being the total number of participants in both experiments) with different patterns of learning rates (see Table S1). The other parameters were set to their mean fitted values across participants and conditions. The results of these analyses confirmed the capacity of our parameter optimization procedure to correctly recover the true parameters in all experimental conditions.

Performance of the CF model

We simulated the CF model in two different environments, where divergence was maximal between buttons ('Environment 1') or between machines ('Environment 2') (see Figure 13A). We tested the performance of the model with three different patterns of factual and counterfactual alpha rates ('subjects', 'flat', 'reverse'). Parameter values were varied according to Table S1, below. For the CF simulations illustrated in Figure 13B, parameter values from the 'subjects' pattern were used, while for the RL simulations, the same parameter values as for the CF model were used ($\alpha_F = 0.5$, $\beta = 0.2$, $\rho_m = 8$, $\rho_b = 5$).

'Environment 1' corresponded to the dependency condition in Experiment 1, whereas 'Environment 2' was similar to the value condition in the same experiment. For the simulations, we used the same task structure as the one experienced by the 16 human participants of our sample, but the model generated its own response, and received the outcome corresponding to this response, on each trial. Each model was simulated 10 times, for each environment and pattern.

Table S1: *Model parameters used in the parameter recovery procedure and to generate model?s performance, as shown on Figures 12C and 13A. Parameters of the model are: the (factual) learning rate α_F, the 3 counterfactual learning rates α_{CF1}, α_{CF2} and α_{CF3}, the reference point P, the exploitation intensity β and the 2 perseveration biases ρ_m and ρ_b. Note that only the values of the counterfactual learning rates differed across the 'subjects', 'flat' and 'reverse' simulations.*

	α_F	α_{CF1}	α_{CF2}	α_{CF3}	P	β	ρ_m	ρ_b
'subjects'	0.5	0.25	0.5	0.5	50	0.2	8	5
'flat'	0.5	0.5	0.5	0.5	50	0.2	8	5
'reverse'	0.5	0.5	0.25	0.25	50	0.2	8	5

Bibliography

Scott, Steven L (2002). "Bayesian methods for hidden Markov models: Recursive computing in the 21st century". In: *Journal of the American Statistical Association* 97.457, pp. 337–351.

4.3 Additional simulations

In this draft, we have presented a novel model for reinforcement-learning, which is emulating a counterfactual outcome to update the unchosen Q-values. While a classical reinforcement-learning model has usually 2 parameters (i.e. the learning rate α and the exploitation intensity β), this counterfactual model has 4 additional parameters:

- the reference point P, used to emulate a counterfactual outcome, and thus a counterfactual prediction error;
- the three counterfactual learning rates α_{CF1}, α_{CF2} and α_{CF3}, associated with each of the unchosen actions, that are weighting the counterfactual prediction error in the unchosen Q-values update.

We will now simulate this novel model in a classical stationary setting (Sutton and Barto, 1998; Cazé and van der Meer, 2013) and show in which conditions and parameter values it outperforms the classical reinforcement learning model.

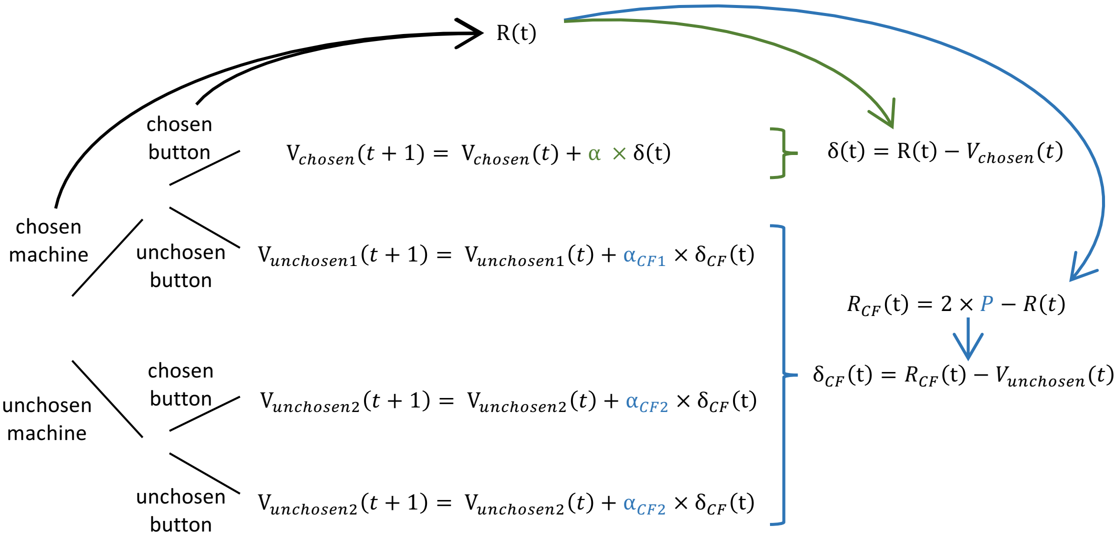

R(t)
chosen button
$V_{chosen}(t+1) = V_{chosen}(t) + \alpha \times \delta(t)$
$\delta(t) = R(t) - V_{chosen}(t)$
chosen machine
unchosen button
$V_{unchosen1}(t+1) = V_{unchosen1}(t) + \alpha_{CF1} \times \delta_{CF}(t)$
$R_{CF}(t) = 2 \times P - R(t)$
$\delta_{CF}(t) = R_{CF}(t) - V_{unchosen}(t)$
unchosen machine
chosen button
$V_{unchosen2}(t+1) = V_{unchosen2}(t) + \alpha_{CF2} \times \delta_{CF}(t)$
unchosen button
$V_{unchosen2}(t+1) = V_{unchosen2}(t) + \alpha_{CF2} \times \delta_{CF}(t)$

Figure 4.1: An illustration of the counterfactual model architecture. On trial t, a machine and a button are chosen, which leads to an obtained reward $R(t)$. Both classical and counterfactual models use the prediction error to update the Q-value of the chosen dyad machine-button. In the counterfactual model, the reward is also used to emulate a counterfactual reward $R_{CF}(t)$ by symmetry with a reference point P. The counterfactual prediction error is weighted by a different counterfactual learning rate α_{CF}, depending on whose unchosen action the Q-value is updated. In green the parameters shared by the classical and counterfactual models, and in blue the parameters specific to the counterfactual model.

4.3.1 The counterfactual learning rates: a theoretical perspective

We have shown that the pattern of counterfactual learning rates is optimal only when adapted to the structure of the task (section "CF model: best-fitting parameters"). We wanted here to understand why. We again used the "dependency" and the "value" conditions from Experiment 1, to be the environments on which we will test the different patterns of learning rates.

Here we assumed that the average of all reward distributions (here 50) was taken as the reference point. It should be noted that the dependency condition has what we call 'button-symmetric' reward distributions. It means that, for a given machine, the outcome received for the chosen button is symmetric to the outcome of the unchosen button. Therefore the counterfactual model should use the counterfactual outcome to update the Q-value of the unchosen button of the chosen machine ($\alpha_{CF1} > 0$), but not the Q-values of the two other unchosen actions ($\alpha_{CF2} = \alpha_{CF3} = 0$).

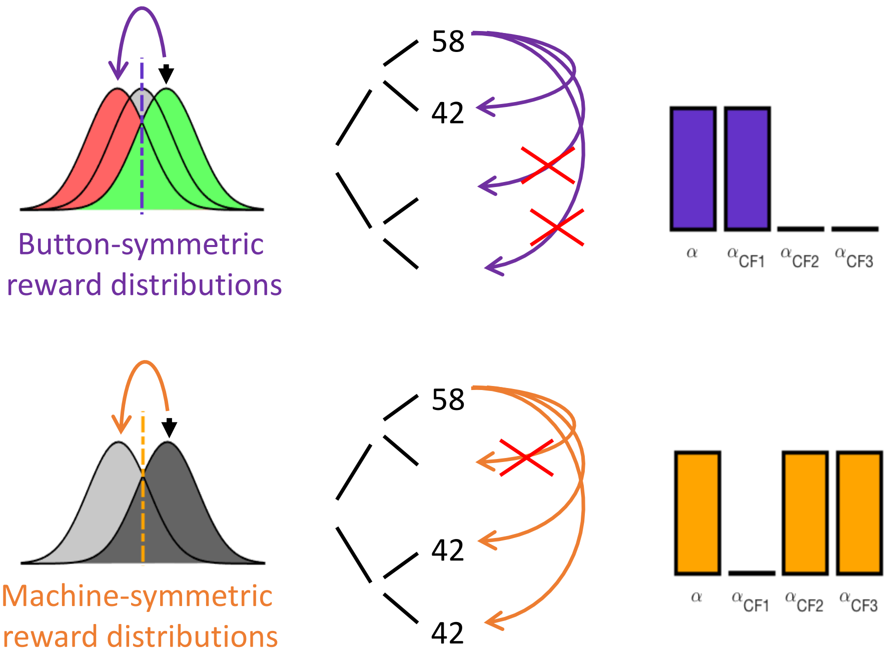

Button-symmetric reward distributions
58
42
α
αCF1
αCF2
αCF3
Machine-symmetric reward distributions
58
42
42
α
αCF1
αCF2
αCF3

Figure 4.2: Left panels: an illustration of the two conditions on which the models will be simulated. Middle panels: the optimal counterfactual updates according to the condition symmetry. Right panels: the pattern of counterfactual learning rates corresponding to these optimal updates. The value of the factual learning rate α is shown here for comparison purposes.

The value condition has the inverse symmetry: its reward distributions are machine-symmetric. There the optimal counterfactual model will update the Q-values of the unchosen machine (for both chosen and unchosen buttons) with the counterfactual outcome ($\alpha_{CF2} > 0$ and $\alpha_{CF3} > 0$), but the Q-value of the unchosen button of the chosen machine should stay unchanged ($\alpha_{CF1} = 0$).

The optimal pattern of counterfactual learning rates can therefore be deduced from the reward distributions of the two conditions.

4.3.2 The counterfactual learning rates: simulations

To test this hypothesis, we ran 10,000 simulations of the two patterns of learning rates, and of a classical reinforcement-learning model. We set the beta value at 0.1 to allow some explorative behavior. As said above the reference point used was 50. The learning rates (factual and counterfactual) were set at 0.1 or 0, according to the pattern tested. The four Q-values are initiated at 50, the average of reward distributions. The simulations lasted for 200 trials in the two first conditions of Experiment 1. Unlike the simulations in the draft, here the action-outcome relations were stationary, i.e., no reversal occurred during the whole task.

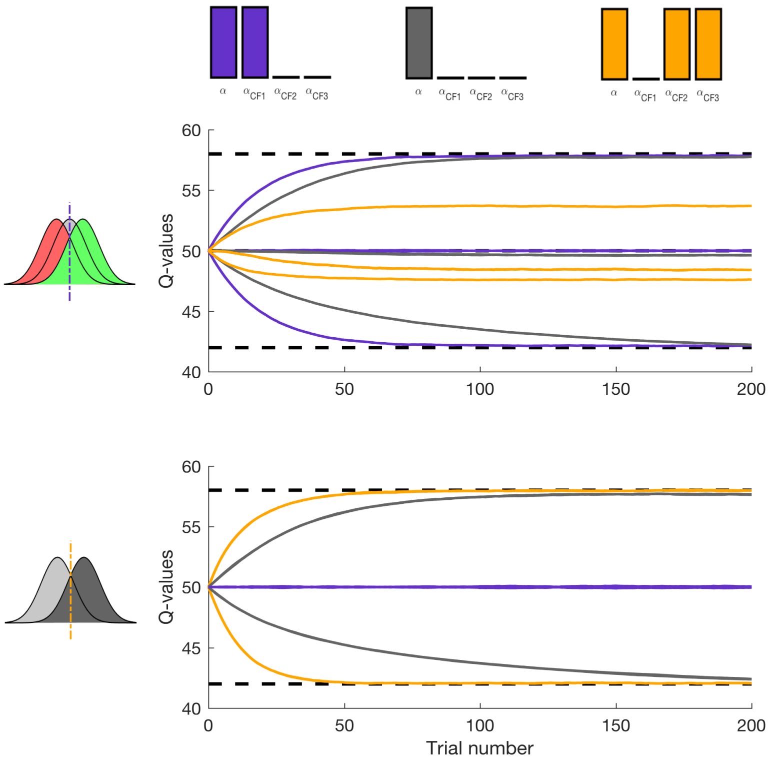


Figure 4.3: The dynamics of the Q-values of counterfactual models with button- (in purple, left panel) and machine-symmetric (in orange, right panel) pattern of counterfactual learning rates. We have also plotted the Q-values dynamics of a classical reinforcement-learning model (in grey) for comparison purpose. The dotted lines represent the average reward for the reward distributions associated with the different actions. In this figure, and for all the following ones, the upper and lower panels show the simulations in the dependency and value conditions respectively.

We looked at the dynamics of the Q-values for the different models in the two environments. In both tasks, the classical Reinforcement-Learning model is approximating well the average reward of each action (58, 50 and 42 for the dependency condition; 58 and 42 for the value condition). Interestingly we can see different dynamics in high and low Q-values: the Q-values approximating 42 were slower to converge than the Q-values approximating 58, as they were associated to actions less frequently chosen by the model.

For the counterfactual model, the Q-values dynamics depended on whether the pattern of counterfactual learning rates was adapted, or not, to the task. For the dependency condition and the button-symmetric counterfactual model, the Q-values not only approximated well the average reward, but also were faster to converge, compared to a Reinforcement-learning model. The same can be observed for the machine-symmetric counterfactual model in the value condition. Interestingly, we can see that for this model, both high and low Q-values displayed the same speed of convergence, as outcomes were used to update both the chosen and unchosen Q-values.

It should be emphasized that the counterfactual models with unadapted patterns of learning rates poorly approximated the underlying average rewards. The machine-symmetric counterfactual model's Q-values converged to biased estimates of average rewards in the dependency condition, while the button-symmetric model was unable to dissociated the high-rewarding actions from the low-rewarding ones (Q-values staying around 50) in the value condition.

The Q-values dynamics were good predictors of a model's performance, as the models are using the difference in Q-values to guide the action selection. We can thus predict that models with well-discriminated Q-values will perform better than models with more similar Q-values, for a fixed exploitation intensity parameter β. Indeed, the counterfactual model whose pattern of learning rates was adapted to the condition outperformed both the Reinforcement-Learning model and the unadapted counterfactual model, in both tasks. It should be noted that counterfactual model was more advantageous than the classical one only when their counterfactual learning rates reflected well the task's structure.

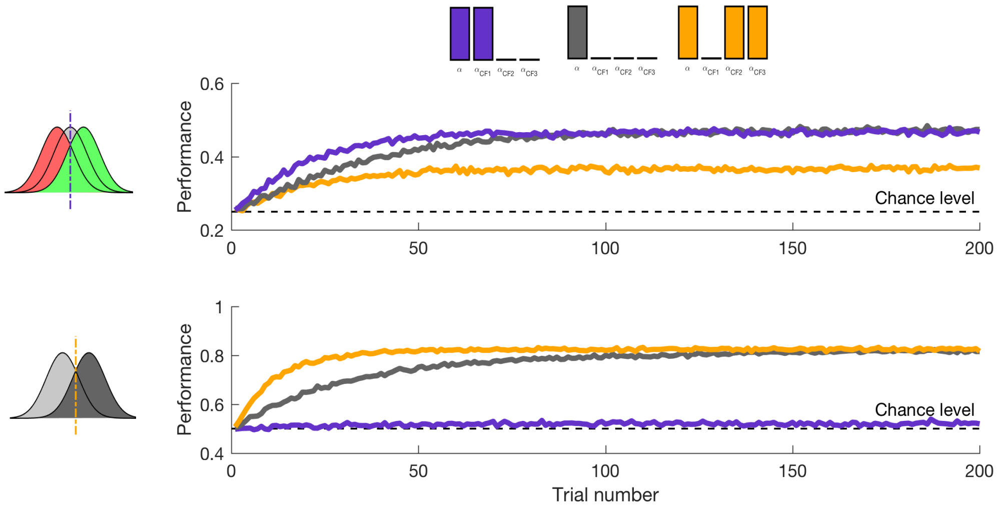


Figure 4.4: The performances of the classical reinforcement-learning model (in grey) and the counterfactual models with button- (in purple) and machine-symmetric (in orange) patterns of counterfactual learning rates. The x-axis represents the trial number. The chance level is represented in dotted line (0.25 for the A-O dependency condition as only one action is correct, but 0.5 for the Outcome value condition, as two actions among 4 possible are associated with the high-rewarding machine).

4.3.3 The reference point: simulations

We then wondered how the dynamics of the counterfactual model may change when the reference point was no longer the underlying average of reward distributions.

We looked at the Q-values dynamics when we set the reference point to the values of 40 and 60 and ran simulations similar to those described above. As we can see, when the reference point departed from the value of 50, the Q-values did not converge anymore to the average reward for each action. They underestimated the real reward average when the reference point was set under 50, and they overestimated them when the reference point was set at 60.

It seemed that the lower the reference point, the higher the difference between Q-values was. We could therefore predict that a reference point lower than 50 would increase the model's performances. Still, we considered purposeless to further ex-

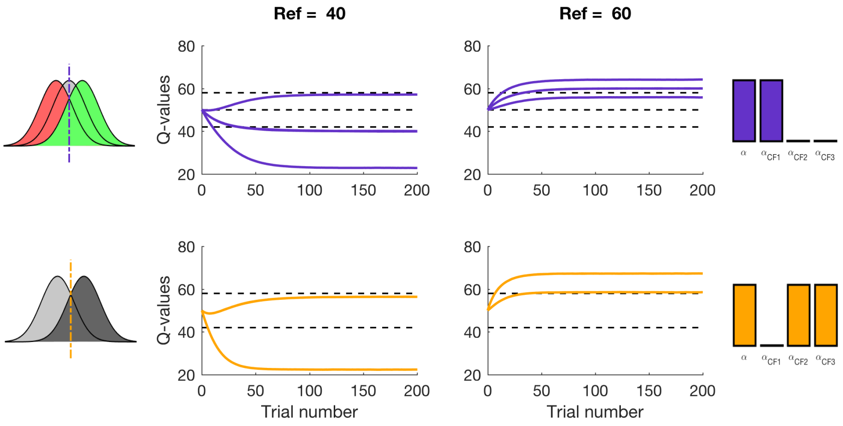


Figure 4.5: The dynamics of the Q-values of the counterfactual models for a reference point P of 40 and 60. We only plotted the model that was shown to be more adapted to the task, therefore the button-symmetric model for the dependency condition, and the machine-symmetric model for the value condition. The dotted lines represent the average reward for the reward distributions associated with the different actions.

plore the mechanisms of this counterfactual model, as it has become a degenerate model, whose internal variables were no longer related to the true statistics of the task.

In summary, when the pattern of counterfactual learning rates was reflecting the task's symmetry, the counterfactual model outperformed the classical reinforcement-learning model. Still a non-adapted counterfactual model performed worse than the classical one, or even at chance. It should be added that the performances of the counterfactual model were reference-dependent. The counterfactual model developed here had never been described before in the reinforcement-learning literature. Our simulations confirmed that this model can be advantageous in stationary symmetrical tasks with more than 2 possible actions.

Chapter 5

Study II

5.1 Introduction

A fundamental experience of everyday life is the feeling that we control our own actions. When these actions produce effects in the environment, we feel that we cause those too. This cluster of experiences is referred to as the Sense of Agency (Haggard and Chambon, 2012).

In the previous study, we were interested in how human participants would compute an on-the-fly estimate of instrumental control in an instrumental conditioning experiment. Here for the first time, we measured participants' feeling of agency during an instrumental task. We used the intentional binding paradigm, as there is compelling evidence supporting a link between intentional binding and sense of agency (Moore and Obhi, 2012).

Sense of agency or subjective control depends on the ability to learn and make use of action-outcome contingencies and one of the more classical algorithm to model this learning originates in the field of reinforcement learning. Our goal was to study the possible correlations between implicit feelings of agency and reinforcement learning processes.

5.2 Our published research article

Try and try again: Post-error boost of an implicit measure of agency

Steven di Costa[1]*, Héloïse Théro[2]*, Valérian Chambon[3] and Patrick Haggard[1]

* co-first authors

[1] *Institute of Cognitive Neuroscience, University College London, London, UK*

[2] *Laboratoire de Neurosciences Cognitives, INSERM-ENS, Département d'Etudes Cognitives, PSL Research University, Paris, France*

[3] *Institut Jean Nicod (ENS-EHESS-CNRS), Département d'Etudes Cognitives, PSL Research University, Paris, France*

Corresponding author: Steven Di Costa (Email: stevendicosta@gmail.com)
Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AZ, UK.

Accepted in Quarterly Journal of Experimental Psychology:
31 May 2017

The sense of agency refers to the feeling that we control our actions and, through them, effects in the outside world. Reinforcement learning provides an important theoretical framework for understanding why people choose to make particular actions. Few previous studies have considered how reinforcement and learning might influence the subjective experience of agency over actions and outcomes. In two experiments, participants chose between two action alternatives, which differed in reward probability. Occasional reversals of action-reward mapping required participants to monitor outcomes and adjust action selection processing accordingly. We measured shifts in the perceived times of actions and subsequent outcomes ('intentional binding') as an implicit proxy for sense of agency. In the first experiment, negative outcomes showed stronger binding towards the preceding action, compared to positive outcomes. Furthermore, negative outcomes were followed by increased binding of actions towards their outcome on the following trial. Experiment 2 replicated this post-error boost in action binding and showed that it only occurred when people could learn from their errors to improve action choices. We modelled the post-error boost using an established quantitative model of reinforcement learning. The post-error boost in action binding correlated positively with participants' tendency to learn more from negative outcomes than from positive outcomes. Our results suggest a novel relation between sense of agency and reinforcement learning, in which sense of agency is increased when negative outcomes trigger adaptive changes in subsequent action selection processing.

Keywords

Agency; learning; intentional binding; time perception; decision-making; motor control

Introduction

Achieving one's goals requires detection of errors and consequent adjustments to behaviour (Balleine and Dickinson, 1998). A distinctive subjective experience accompanies committing an error and registering its outcome (Charles, King, and Dehaene, 2014). Sense of agency is defined as the feeling of controlling one's actions and their effects in the outside world (Haggard and Chambon, 2012). However, the extensive literature on learning from errors (Dayan and Niv, 2008) has evolved largely independently from the literature on sense of agency. Therefore, in two experiments, we investigated how errors in a reversal-learning task influence sense of agency.

Explicit judgements of control or agency are influenced both by performance bias (Metcalfe and Greene, 2007) and by a general self-serving bias (Bandura, 1989). A confounding effect of errors on explicit agency judgements therefore seems inevitable. The intentional binding paradigm (Haggard, Clark, and Kalogeras, 2002; for a review, see Moore and Obhi, 2012) offers an implicit measure related to sense of agency, which may be less subject to task demand characteristics. Participants report the time of an action or of its outcome. If the outcome follows the action with a short and constant latency, the perceived time of the action tends to shift towards the subsequent outcome. Similarly, the perceived time of the outcome tends to shift towards the preceding action. Critically, these effects are stronger for voluntary actions than for involuntary movements (Haggard, Clark, and Kalogeras, 2002). Intentional binding may be one instance of a more general temporal binding effect that applies to causal relations (Buehner and Humphreys, 2009; but see Cravo, Claessens, and Baldo, 2009; Cravo, Claessens, and Baldo, 2011). However, experimental designs that contrast appropriately chosen conditions can nevertheless use binding measures as a proxy measure to investigate different components of sense of agency.

Previous laboratory research on sense of agency often lacked ecological validity. For example, intentional binding studies have investigated associations between a single action and a single outcome without any significance or value for the participant (Haggard, Clark, and Kalogeras, 2002). Outside the laboratory, however, actions are embedded in a rich perceptual, affective and social landscape. People frequently select one action from several possible in a given situation, to achieve a desired goal. Only a few studies have attempted to link implicit measures of sense of agency with outcome valence. In Takahata et al. (2012), participants' actions caused tones that were associated with monetary rewards or penalties. They found reduced binding for penalty trials compared to neutral or rewarded trials. Yoshie and Haggard (2013) used human vocalizations as either negative or positive action outcomes. They found that negative vocalization outcomes were associated with a reduction in binding compared to neutral and positive vocalization outcomes. Neither study manipulated the effects of contingency between participants' actions and the rewards received, and neither study tried to distinguish the informational value of outcomes from their reward value. In the present work, we manipulated occurrence of rewards to investigate effects of reinforcement and learning.

Accordingly, we have combined intentional binding with reward-based decision-making, seemingly for the first time. We used a probabilistic reversal-learning approach (Cools et al., 2002; Rolls, 2000), which requires the participant to continuously learn action-outcome mappings, and update their action choices according to error feedback. The action-outcome structure of reversal learning can be combined straightforwardly with the intentional binding paradigm. Furthermore, probabilistic reversal learning can be challenging enough to require consistent cognitive engagement. In contrast, humans often readily achieve agency in situations involving new stable action-outcome relations, so instrumental learning and sense of agency emerge too rapidly to be measured with current paradigms.

In reversal learning, participants need to monitor the outcome linked to each action and then correctly update their expectations so as to select their next action accordingly (Sutton and Barto, 1998). A central issue in research on learning is how behaviour changes trial by trial in response to feedback (Daw, 2011). In this study, we were interested in the fluctuation of sense of agency that accompanies reward-based decision-making. We predicted that the occurrence of rewards might influence not only the intentional binding associated with a given outcome but also the intentional binding reported on the subsequent trial.

Experiment 1: method

Participants

This study was approved by the UCL Research Ethics Committee and conformed to the Declaration of Helsinki. In the absence of any previous study combining intentional binding with reward-guided decision-making, the sample size was based on a study of intentional binding with valenced trial outcomes (Yoshie and Haggard, 2013). A total of 16 participants (nine females, all right-handed, mean age = 23 years, age range = 18-41 years) completed the experiment and were paid £8/hr plus a bonus for correct responses. Data from one participant were lost due to a technical error. All participants reported normal or corrected-to-normal vision and hearing.

Procedure

Participants were seated at a standard computer keyboard and screen. They fixed a clock with a single rotating hand. The clock diameter was 20mm and the hand completed one full rotation within 2560ms. In baseline conditions, participants pressed a key at a time of their free choice or heard an auditory tone at a random time. In 'agency' conditions, participants both pressed a key and heard a tone. The tone occurred 250 ms after the key press. Participants were instructed to wait for one full rotation of the clock before pressing a key. Tones were either high (2000 Hz) or low (500Hz) in frequency and lasted 100ms. The high tone

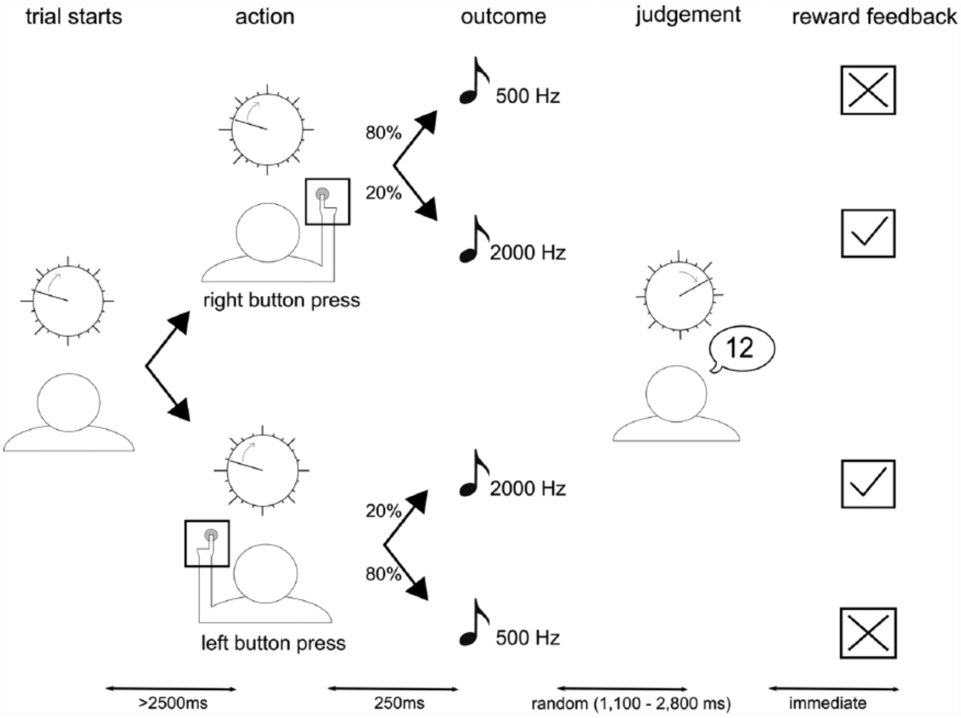


Figure 1: *Timeline of events on a typical agency trial in Experiment 1. The trial started when the clock hand began to rotate. At a time of their free choice, allowing for at least one full rotation of the clock hand, participants pressed one of two keys. Key presses were followed after 250 ms by a high or low frequency tone. The clock hand continued to rotate for a random interval and then stopped. Participants then reported the time they perceived the action or tone to have occurred. Immediate visual feedback then confirmed the earlier auditory signal, indicating a reward (tick) or non-reward (cross).*

was always the 'correct' tone and was associated with a monetary reward. Informal piloting indicated that participants had clear prior associations, interpreting high tones as positive and low as negative. These may reflect common conventions of everyday electronic devices. Therefore, we did not counterbalance the tones across participants. The 'F' and 'J' keys of a standard keyboard were used for left- and right-hand responses.

Following the tone (or the key press if no tone), the clock hand continued to rotate for a random interval between 1100 and 2800ms and then disappeared. Participants then used the keyboard to report the time that they pressed the button or the time that they heard the tone, according to condition (Figure 1).

Baseline action and tone measures were first taken in six separate blocks of 20 trials (see below) in pseudorandom order, to provide estimates of the perceived time of each action and each tone when presented alone, and when presented as the only event within a block, or mixed with the alternative action or tone. Next, participants completed two counterbalanced 'agency' blocks. In one block, they reported the perceived time of the action, in the other block the perceived time of the tone. Finally, the six baseline conditions were repeated in the reverse order. Thus, there were always 40 trials in each condition, and conditions were always blocked.

In the agency conditions, one key delivered rewarded high tones with a probability of 0.8 and the other key with probability of 0.2. The mapping was maintained across a run of several trials, until the participant had selected the key that produced the high tone (i.e., the reward) between five and seven times consecutively (randomized). Probability mappings then reversed. Nine such reversals occurred in each block, so each block involved 10 'runs' of responses. The actual number of key presses per block therefore depended on how rapidly each participant learned the 'correct' key.

The cumulative total of rewarded trials was displayed at the end of each trial. At the end of each block, all participants were told they had reached the threshold number of rewarded trials required to trigger a bonus. In fact, this threshold was fictitious, and a bonus of £3 for each block was paid at the end of the experiment. This arrangement ensured that participants were not overpaid for prolonging the experiment by repeatedly making incorrect responses.

In each trial, a visual feedback indicating either reward (tick) or no reward (cross) reward was presented for 1s after each judgement, followed by an inter-trial interval of 1 s. The visual signal recapitulated the information previously conveyed by the auditory tone, but was included to facilitate decision-making on the next trial, without placing strong demands on memory.

We did not directly probe participants' awareness of action-outcome contingencies. Rather, we considered

Table 1: *Mean (M) and standard deviation (SD) of judgement errors (ms) in baseline and agency conditions in Experiment 1.*

	Baseline before		Baseline after	
	M	SD	M	SD
Action (left hand)	−42	87	22	60
Action (right hand)	−40	63	−17	88
Action (free choice)	−40	103	−16	78
Tone (high)	15	70	51	72
Tone (low)	25	76	29	68
Tone (mixed)	12	79	29	84

	All agency trials	
	M	SD
Action	42	64
Tone	−83	135

that generating a sequence of repeated key presses of the 'good' key, and thus triggering a reversal, was a sufficient indicator of learning. All stimuli were presented using LabView 2012 (National Instruments, Austin, TX).

Baseline measures

Baseline judgement errors are presented in Table 1.

No significant differences were observed between the baseline blocks in the perceived times of key presses in milliseconds for left- and right-hand responses ($F_{1,14} = 0.176, p = 0.681, \eta_p^2 = 0.012$), mixed or repeated presentation ($F_{1,14} = 0.236, p = 0.635, \eta_p^2 = 0.017$), or for pre- or post-experiment blocks measures ($F_{1,14} = 3.137, p = 0.098, \eta_p^2 = 0.183$).

Consequently, all action baseline blocks were collapsed in further analysis. Likewise, no significant differences were observed in the perceived times of high- and low- frequency auditory tones ($F_{1,14} = 0.599, p = 0.452, \eta_p^2 = 0.041$), for mixed or repeated presentation ($F_{1,14} = 1.827, p = 0.198, \eta_p^2 = 0.115$) or for pre- or post-test measures ($F_{1,14} = 3.107, p = 0.1, \eta_p^2 = 0.182$). Consequently, these were also collapsed in further analysis.

Analysis

Perceptual shifts were then calculated for each participant and each condition by subtracting the relevant mean baseline error for action or tone from that in agency trials. A positive action binding measure therefore corresponds to a shift of the perceived time of the action towards its outcome and a negative outcome binding measure to a shift of the perceived time of the outcome towards the action. Agency trials were categorized according to two design factors:

1. whether the outcome received on the current trial was rewarded (high tone) or not rewarded (low tone)
2. whether feedback on the *previous* trial was rewarded or not rewarded.

Experiment 1: results

The overall ratio of trials with non-rewarded outcomes to rewarded outcomes was 0.6:1 (mean number of trials per block = 109, standard deviation [SD] = 35).

Performance

Participants learned the action-outcome contingencies (Figure 4a). As the criterion for advancement was set at five to seven presses of the more rewarded key, participants' performances were necessarily 100% before reversal of action?outcome mappings. Reversal events unsurprisingly triggered errors. We analysed the proportion of correct choices using a repeated-measure analysis of variance (ANOVA) with trial number after reversal as a factor. The trial number had a significant effect on participants' performance ($F_{4,56} = 66.2, p < 0.001, \eta_p^2 = 0.250$). As the figure shows, participants adapted their responses after a few reversal-induced errors occurred.

Intentional binding

Action and outcome binding data are shown in Figure 2. Action binding data were subjected to a 2 × 2 ANOVA with factors of current trial outcome: low tone (no reward) or high tone (reward) and previous trial outcome. There was a highly significant effect of previous trial outcome (low tone: M = 87.2, SD = 62.8; high tone: M = 63.0, SD = 49.2), with stronger action binding following low tones than following high tones ($F_{1,14} = 9.20, p = 0.009, \eta_p^2 = 0.397$). There was no effect of current trial outcome (low tone: M = 69.8, SD = 60.2; high tone: M = 74.2, SD = 48.1; $F_{1,14} = 1.72, p = 0.210, \eta_p^2 = 0.110$) and no interaction ($F_{1,14} = 0.01, p = 0.941, \eta_p^2 = 0.000$).

A similar ANOVA was performed for outcome binding. This showed a significant effect of current trial outcome (low tone: M = -119.3, SD = 100.4; high tone: M = -105.1, SD = 93.9), with low tones being more strongly bound towards actions than high tones ($F_{1,14} = 6.32, p = 0.025, \eta_p^2 = 0.311$). There was no effect of previous trial outcome (low tone: M = -114.6, SD = 93.6; high tone: M = -108.2, SD = 99.8; $F_{1,14} = 0.02, p = 0.89, \eta_p^2 = 0.002$) and no interaction ($F_{1,14} = 1.89, p = 0.19, \eta_p^2 = 0.119$).

Experiment 1: discussion

In a reversal-learning task, we observed that non-rewarded outcomes were more strongly bound back to their actions than rewarded outcomes. Our results

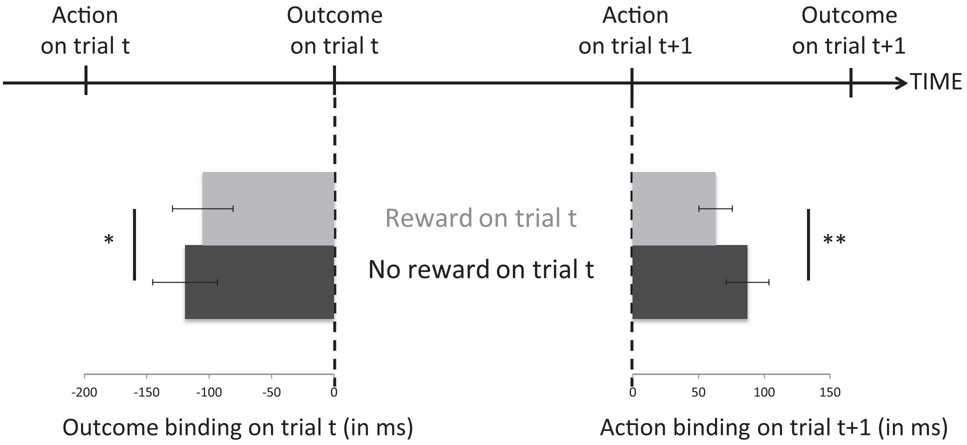


Figure 2: *Outcome and action binding in Experiment 1. (Error bars represent standard errors.)*

therefore differ markedly from previous studies of binding and valence (Takahata et al., 2012; Yoshie and Haggard, 2013), in which negative outcomes showed less binding than positive outcomes. This difference may reflect the presence of both error-based learning and action selection in reward-based decision-making in the current task, but not in those previous studies.

Furthermore, action binding on the trial *following* a non-rewarded outcome was stronger than following a rewarded outcome. To our knowledge, this is a first time that previous trial outcome has been reported to have a *sequential* effect on action binding. Some previous studies reported effects of the occurrence (Moore and Haggard, 2008) or timing (Walsh and Haggard, 2013) of preceding outcomes on subsequent action binding, but those studies did not involve the crucial element of selection between alternative outcomes. In our study, unlike previous work, the sequential effects on action binding may be linked to errors and to learning.

In a second experiment, we therefore aimed to replicate this post-error boost of action binding and investigate whether it was indeed dependent on learning and reward. We thus added a 'non-learning' condition in which participants made actions and received outcomes as before, but action-outcome mappings were now entirely unpredictable. We explicitly informed participants about the nature of these two conditions. We predicted stronger action binding in the learning condition than in the random condition.

Experiment 2: method

Participants

A total of 30 participants (21 females, all right-handed, mean age = 28 years, age range = 21-53 years) completed the experiment and were paid £7.5/hr plus a bonus for correct responses and precision. The number of participants was increased, compared to Experiment 1, to allow us to correlate intentional binding measures with learning measures across participants.

General procedure

The general procedure was similar to Experiment 1, except for the following: here the keys used to select an action were the 'right-arrow' and 'left-arrow' keys of a standard keyboard, using the index and middle fingers of the right hand, respectively. Participants reported the time by typing on the keyboard with their left hand. No visual feedback was presented following timing judgements. Participant reports from Experiment 1 indicated that they did not particularly attend to the visual feedback. Because it was redundant with the tone frequency, it was omitted in Experiment 2.

We focused on measuring action binding, and not tone binding, because action binding has been linked to outcome prediction mechanisms (Engbert and Wohlschläger, 2007) and to experience-dependent plasticity (Moore and Haggard, 2008). Furthermore, excluding tone binding allowed us to increase the trial numbers in agency blocks without making the experiment excessively long.

Agency conditions

Besides the baseline measures, participants completed five blocks of 30 trials in the learning condition, and five in the random condition, in pseudo-randomized order. In the learning condition, one key delivered rewarded high tones with a probability of 0.8 and the other key with probability of 0.2. The high tone was always the 'correct' tone, and participants were told to learn which key was most frequently associated with the high tone. We also explicitly informed subjects that reversals of the action-tone mapping would occur occasionally and unpredictably. These explicit instructions aimed to reduce the high inter-individual variability in performance found in Experiment 1, by clarifying the task for poorer performers. Furthermore, reversals now occurred after a variable number of trials (randomly 6, 10 or 14 trials) so participants could not predict when they would occur. We adjusted the run length after the last reversal in the block to ensure the same number of trials for each participant. At the end of each block of the learning condition, if participants achieved a threshold of at least 20 rewarded trials, they gained a bonus of 50p. We used a large blockwise reward rather than smaller trialwise rewards, to avoid satiety after several successful trials and to maintain motivation throughout.

In the random condition, the probability of hearing a high tone or a low tone was unrelated to the key chosen (50%/50%). Participants were explicitly told that their choice of action would not influence the tones they would hear. In the learning condition, they were instructed to 'find the good key, maximizing the number of high tones', while in the random condition they were told, 'whichever action is chosen, it will have no influence on the following tone'. Since learning could not be used to maximize reward in this condition, the number of high-tone trials did not lead to a monetary bonus. This arrangement ensured that participants were not incentivized to search for contingencies that did not exist. Although this creates a motivational difference between the two conditions, this bias is intrinsic to any reinforcement-learning experiment (O'Doherty, 2014). Furthermore, at the beginning of each block, participants were explicitly told which condition they were in.

As before, participants reported the timing of their action. To further improve the precision of our measure, we instructed participants that at the end of each block they would receive an additional 25p if they improved the precision of timing estimates relative to the previous block. We used the SD of their judgement errors to measure precision – note that this measure is independent of the mean timing judgement and thus independent of action binding estimates. Thus, in the learning condition, participants were rewarded for precision of timing judgements and for choosing the 'correct' key. In the random condition, they were rewarded only for precision of timing judgements.

Baseline measures

We also measured the perceived times of actions presented without tones in a baseline condition. Participants performed two baseline blocks of 20 trials each, at the beginning and end of the agency session. In baseline blocks, participants freely chose which of the two keys to press. Baseline judgement errors are presented in Table 2.

No significant differences were observed in the perceived times of key presses in milliseconds for left- and right-hand responses ($F_{1,29} = 1.01, p = 0.319, \eta_p^2 = 0.018$) or for pre- or post-experiment blocks measures ($F_{1,29} = 0.129, p = 0.721, \eta_p^2 = 0.002$). Consequently, action baseline blocks were collapsed in further analysis.

Table 2: *Mean (M) and standard deviation (SD) of judgement errors (ms) in baseline and agency conditions in Experiment 2.*

	Baseline before		Baseline after	
	M	SD	M	SD
Action (free choice, left hand)	-27	139	-10	112
Action (free choice, right hand)	-34	77	-47	105

	All agency trials	
	M	SD
Action (learning condition)	-5	110
Action (random condition)	-28	93

Analysis

Action binding was calculated for each participant and each condition by subtracting the relevant mean baseline error from the error in agency trials. Agency trials were categorized according to three design factors:

1. whether the outcome on a given trial was a high or low frequency tone (associated with a positive or negative outcome, respectively, in the learning condition);
2. whether the trial was in the learning or random condition;
3. whether the outcome on the *previous* trial was a high or low frequency tone.

Action binding data were then subjected to a 2 × 2 × 2 ANOVA.

Experiment 2: results

Performance

In the learning condition, participants demonstrated an ability to learn the correct action. As in Experiment 1, the trial number after reversal had a significant effect on participants' proportion of correct choice ($F_{5,145} = 57.14, p < 0.001, \eta_p^2 = 0.200$). They quickly returned to initial performance levels after a reversal event (Figure 4a).

Action binding

Action binding data are shown in Figure 3. A 2 × 2 × 2 ANOVA revealed a highly significant main effect of condition (learning condition: M = 28.8, SD = 53.3;

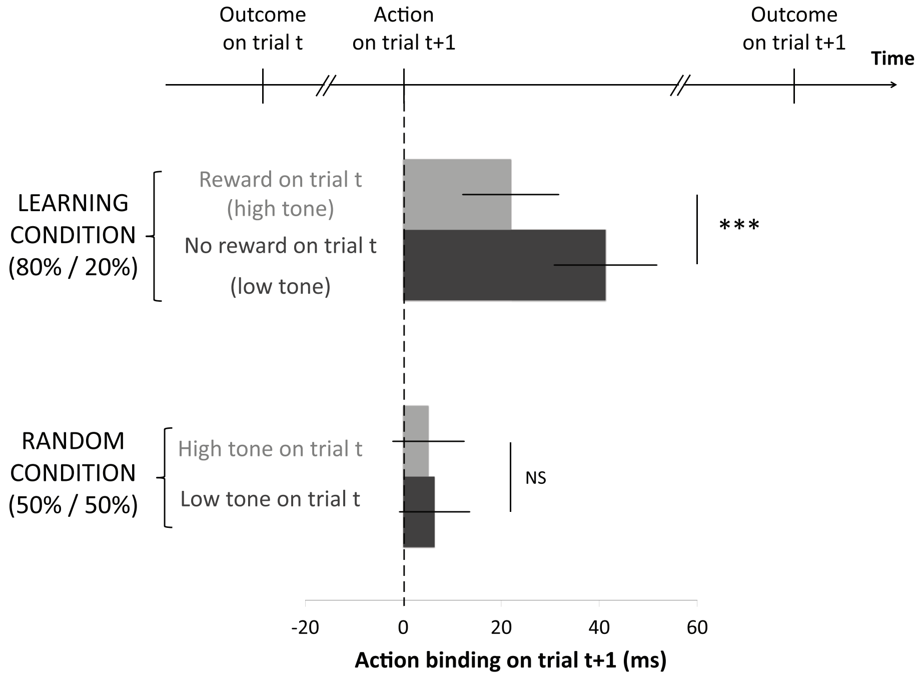


Figure 3: *Mean action binding (ms) following a rewarded (light grey) or non-rewarded (dark grey) outcome on the previous trial, for both random and learning conditions. Note that the high/low tones were associated with rewarded/non-rewarded outcome in the learning condition, but not in the random condition. (***: $p < 0.001$)*

random condition: M = 5.6, SD = 39.0), with stronger action binding in the learning condition compared to the random condition ($F_{1,29} = 17.48, p < 0.001, \eta_p^2 = 0.376$). There was no effect of current trial outcome (low tone: M = 16.4, SD = 42.7; high tone: M = 17.8, SD = 44.7; $F_{1,29} = 0.02, p = 0.896, \eta_p^2 = 0.001$). Importantly, we found a significant main effect of previous trial outcome (low tone: M = 21.3, SD = 43.2; high tone: M = 14.7, SD = 44.7; $F_{1,29} = 14.56, p < 0.001, \eta_p^2 = 0.334$) and also a highly significant interaction between learning condition and previous trial outcome ($F_{1,29} = 9.71, p = 0.004, \eta_p^2 = 0.251$; see Figure 3).

We performed *simple-effect t-tests* to further investigate this interaction. In the learning condition, non-rewarded outcomes significantly increased the action binding on the following trial compared to rewarded outcomes (simple-effect paired t-test: $t_{29} = 3.73, p < 0.001, Cohen?sd = 685$). This difference was numerically almost abolished and became statistically non-significant, in the random condition ($t_{29} = 0.46, p = 0.646$; see Figure 3).

No other interactions were significant (current trial outcome × condition: $F_{1,29} = 0.33, p = 0.573, \eta_p^2 = 0.011$; current trial outcome × previous trial outcome: $F_{1,29} = 1.01, p = 0.323, \eta_p^2 = 0.034$; and current trial outcome × condition × previous trial outcome:$F_{1,29} = 0.13, p = 0.718, \eta_p^2 = 0.005$).

Experiment 2: discussion

With some changes in implementation, we replicated the post-error boost in action binding in the learning condition. Crucially, we showed that this effect is *specific* to a learning context and is absent when participants cannot learn stable action-outcome relations. Our results therefore provide strong evidence that action binding reflects the ability to influence events through learning to improve one's own action choices. Critically, this learning depends on previous error feedback.

We next used a formal reinforcement-learning model to investigate how the post-error boost in action binding is related to how people learn to maximize rewards. Reinforcement-learning models distinguish between the learning opportunities offered by errors and by rewards, respectively. Interestingly, these two elements of learning are differentially expressed across the population. Negative learners are better at avoiding negative outcomes, while positive learners are better at choosing positive outcomes. Interestingly, the electroencephalogram (EEG) feedback-related negativity (FRN) evoked by an error signal has been found to be larger in negative learners than in positive learners (Frank, Woroch, and Curran, 2005). Similarly, we hypothesized that the post-error boost in action binding might be positively correlated with participants' bias to learn more from negative than from positive outcomes.

Statistical modelling of results from Experiments 1 and 2

Method

We fitted an established model of reinforcement learning to investigate whether inter-individual variance in asymmetric learning is correlated with the post-error boost in action binding. According to the reinforcement-learning algorithm, each of the two possible actions (choosing the left or right button) was associated with an internal value called an action value (Sutton and Barto, 1998). The values themselves are hidden but are thought to drive choices between alternative actions.

Value updating. The model is based on the concept of prediction error, which measures the discrepancy between actual outcome value and the expected outcome for the chosen action (i.e., the chosen action value):

$$\delta(t) = Outcome(t) - Value_{Chosen}(t)$$

Prediction error is then used to update the value of the chosen action. The values were set as 0.5 at the beginning of each block. Because we were interested in the specific effect of rewarded and non-rewarded outcomes, we set two different learning rates, α^+ and α^-, to reflect different updating processes after a positive or negative prediction error (Lefebvre et al., 2016; Niv et al., 2012). This asymmetrical model therefore accounts for individual differences in the way participants learn from positive and negative outcomes.

$$Value_{Chosen}(t+1) = \\ Value_{Chosen}(t) + \begin{cases} \alpha^+ \times \delta(t) \text{ if } \delta(t) > 0 \\ \alpha^- \times \delta(t) \text{ else} \end{cases}$$

We then normalized the action values of the two possible actions by keeping their sum constant.

We also constructed a reduced model with only one learning rate for both rewarded and non-rewarded outcomes, and the Aikake Integration Factor (AIC) comparison showed that the AIC of the two learning rate model was significantly lower than the AIC of the one learning rate model for Experiment 1 (paired t-test : $t_{14} = 4.56, p < 0.001$) and for Experiment 2 ($t_{29} = 2.37, p = 0.025$). The model with two learning rates (α^+ and α^-) was thus the best fitting model.

Decision rule. In the model, the action with the higher action value is more likely to be selected. The probability to choose an action will depend on the two action values and on the 'inverse temperature' parameter β, which represents the strength of the action values' effect on action selection:

$$P_{ChoosingLeft} = \frac{e^{\beta \times Value_{Left}}}{e^{\beta \times Value_{Left}} + e^{\beta \times Value_{Right}}}$$

Parameter fitting and simulations. We fitted the model parameters based on participants' choices on each trial. The three parameters fitted were the two learning rates, α^+ and α^-, and the inverse temperature β. They were fitted independently for each participant, on the data from the learning condition in Experiments 1 and 2. The best parameters chosen were those that maximized log likelihood (LLH), defined as the sum of the log of the model's fit to participant's action choices. Thus, LLH values close to 0 indicate a good model fit. To test the different possible combinations of parameters, we used a slice sampling procedure (Bishop, 2006). More precisely, using three different starting points drawn from uniform distributions for each parameter, we performed 10,000 iterations of a gradient ascent algorithm to converge on the set of three parameters that best fitted the data.

Results

From the fitted parameters, we simulated the model's choices and found a generally good match with participants' behaviour (Figure 4a). The probability of model selecting the same action as the participant was M = 0.80, SD = 0.06 in Experiment 1; and M = 0.83, SD = 0.09 in Experiment 2. Thus, our reinforcement-learning model seemed to accurately reflect participants' learning processes. Similar to Lefebvre et al. (2016), we found overall higher learning rates for rewarded outcomes than for non-rewarded outcomes (Experiment 1: α^+: M = 0.89, SD = 0.13 and α^- : M = 0.48 SD = 0.14; $t_{14} = 9.15, p < 0.001$ and Experiment 2: α^+: M = 0.67, SD = 0.27 and α^-: M = 0.51 SD = 0.23; $t_{29} = 3.26, p = 0.003$), justifying the use of an asymmetrical model.

We further calculated the normalized learning rate asymmetry (Lefebvre et al., 2016; Niv et al., 2012), defined as:

$$\frac{\alpha^- - \alpha^+}{\alpha^- + \alpha^+}$$

to investigate whether the post-error agency boost could be related to the outcome-specific learning rate. We defined our post-error boost in action binding as the difference between action binding after a non-rewarded outcome and action binding after a rewarded outcome, as before. For Experiment 1, we did not find any relation between post-error agency boost and normalized learning rate asymmetry ($t_{13} = -0.66, p = 0.518, R^2 = 0.03$). However, we found a positive correlation between post-error agency boost and normalized learning rate asymmetry in the learning condition of

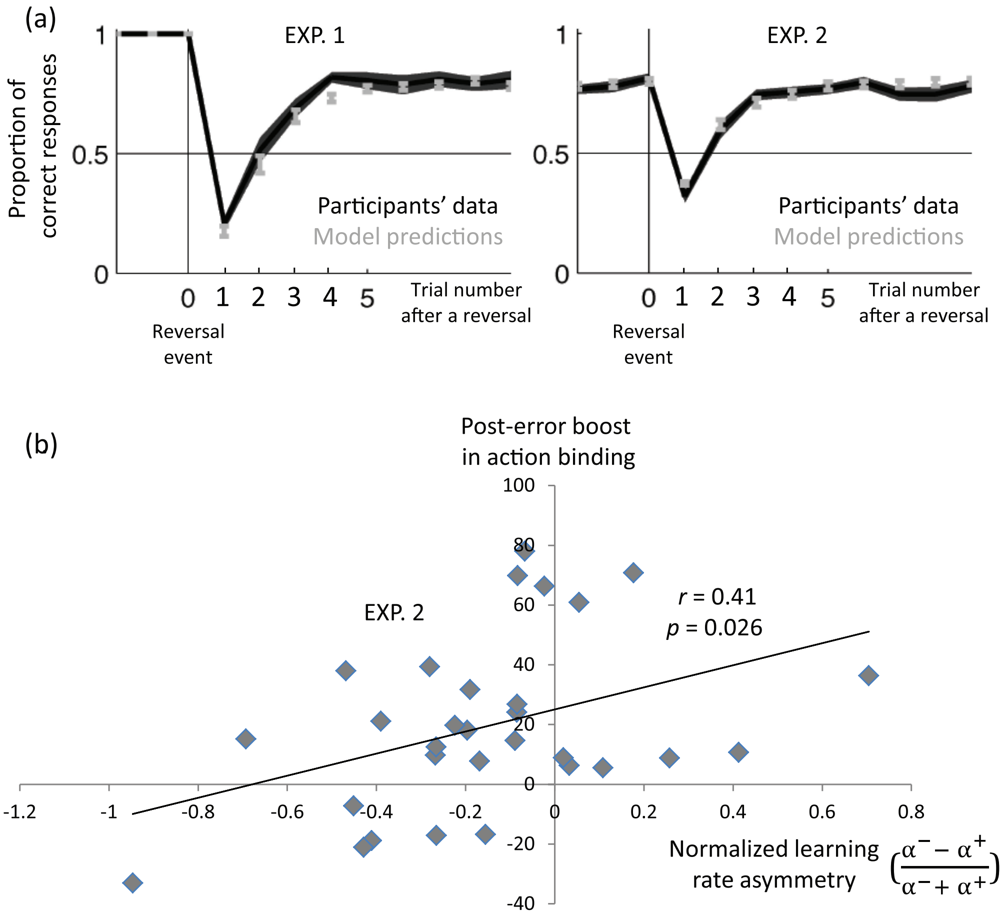


Figure 4: *(a) Proportion of correct responses before and after a reversal event for Experiment 1 (left panel) and Experiment 2 (right panel). Participants' data are in black and predictions of the reinforcement-learning model are in grey. (b) Post-error boost in action binding plotted against the normalized learning rate asymmetry for Experiment 2.*

Experiment 2 ($t_{28} = 5.6, p = 0.026, R^2 = 0.17$; Figure 4b), implying that individuals who learn from errors also show a strong post-error agency boost. The absence of any effect in Experiment 1 may reflect the lower statistical power and may also reflect the very restricted inter-individual variability in learning rate asymmetry (asymmetry in Experiment 1: M = -0.31, SD = 0.14 and in Experiment 2: M = -0.15, SD = 0.32; F-test for comparison of sample variances: $F_{29,14} = 5.29, p = 0.002$).

Finally, we explored whether other confounding factors, in addition to normalized learning rate asymmetry, could predict individual variability in post-error agency boost in Experiment 2. In particular, an alternative view hypothesizes that the post-error agency boost could merely reflect saliency of rare error events, akin to the non-specific alerting effect of an oddball, rather than any relation between errors and learning. This alternative model also predicts a negative relation between an individual's post-error agency boost and the frequency of their errors, yet no such relation was found ($t_{28} = 0.53, p = 0.603, R^2 < 0.001$), and the sign was not as predicted.

General discussion

We have shown that intentional binding, the compression of the temporal interval between an action and its outcome, is sensitive to the occurrence of rewards in a reinforcement-learning environment. Intentional binding has been proposed as an implicit measure of sense of agency (Moore and Obhi, 2012). The capacity to choose between actions in order to obtain desired outcomes seems essential for functional control of actions in everyday life – indeed, this is the standard meaning of the term 'sense of agency' in the social sciences (Haggard, 2017). However, previous experimental studies have not convincingly linked the experience of action to acquiring control over outcomes. Our reversal-learning task forced participants to continuously learn relations between actions and outcomes. Previous studies showed that intentional binding is sensitive to economic (Takahata et al., 2012) and affective (Yoshie and Haggard, 2013) valence, but these studies did not address how outcomes can guide learning and decision-making. Here, we describe for the first time how outcome success or failure influences the sense of agency in a dynamic learning environment.

Experiment 1 found that the tone indicating no reward was more strongly bound back towards the action that caused it than the tone indicating a reward. This effect was small and contrary to previous results (Takahata et al., 2012; Yoshie and Haggard, 2013) so its meaning remains unclear. Those studies suggested that the well-known self-serving bias (Bandura, 1989) might influence not only explicit attributions of agency but also implicit measures of the basic experience of agency. However, our study adds an additional, important element of learning, which those earlier studies lacked. The effects of learning from errors appear to replace or outweigh the effects of valence. In our design, errors provided important evidence for learning what action to perform next. In contrast, the valence of outcomes in previous experiments was completely predictable and unrelated to action choices. Future studies could directly compare these two conditions in the same participants.

We also found stronger action binding following a non-rewarded outcome than following a rewarded outcome, across two studies. To date, only a few studies have considered trial-to-trial variation in intentional binding (Moore and Haggard, 2008; Walsh and Haggard, 2013). Both these studies showed that experience on recent trials can influence binding on subsequent trials. However, neither study involved learning to choose between alternative actions in order to optimize outcomes. Specifically, in neither experiment could participants choose between alternative actions, nor did the outcomes have any value or particular significance for the participant. Experiment 2 replicated this post-error boost in action binding in a new and somewhat larger sample. Experiment 2 further showed that it was absent in a condition where actions and tones were identical, but the action-outcome mapping was random and therefore could not be learned. This specificity allows us to discount purely perceptual effects of high/low tones on subsequent action binding.

The concept of ‘cognitive control’ refers to the control and monitoring of cognitive resources to achieve successful task performance. Errors signal a failure of effective control and trigger a number of adaptations, notably ‘post-error slowing’ (Danielmeier and Ullsperger, 2011). Post-error slowing is classically associated with increased caution in action selection after errors (Dutilh et al., 2012). The relation between post-error agency boost and post-error slowing remains unclear. However, it seems unlikely that a mere transient increase in availability of general cognitive resources devoted to action selection, as suggested by conflict adaptation theories, can explain the increase in post-error action binding. A general boost in attention following an error would be expected to cause a general increase in precision of timing judgements, reducing judgement errors and therefore *reducing* both action binding and tone binding effects – yet we found a specific *increase* in judgement errors for actions only. Instead, we suggest that post-error binding may reflect a specific strategic adaptation to the information value of the trial following an error. This adaptation reflects the fact that errors may be highly informative for future action. For example, following an error in a probabilistic reversal-learning task, it is important to decide whether the action-outcome mapping has changed or not. Was the just-experienced error simply ‘noise’ or does it require a change in behaviour? We suggest that strongly linking actions to outcomes on the trial following an error may be an important element for this classic credit-assignment problem and for guiding future action choices. Taken overall, we suggest that cognitive control mechanisms engaged when people make errors may have two distinct effects: an increase in cognitive resources to restore performance and an increase in the experiential link between action and outcome. The latter effect could trigger a post-error boost in agency. However, our study cannot identify for certain the direction of any causal relation between post-error agency boost and learning from errors.

The computations underlying reinforcement learning are classically thought to take place between the moment when the outcome is received and the moment when the next action needs to be performed (Rangel, Camerer, and Montague, 2008; Sutton and Barto, 1998). During that time, the outcome is used to update participants’ expectancy regarding their available actions. Reinforcement-learning processes are thus thought to correspond to this sequential effect. Therefore, we formally modelled participants’ choices using a reinforcement-learning model. Consistent with the literature, we found that participants learned more from rewarded than from non-rewarded outcomes (Lefebvre et al., 2016; Niv et al., 2012). This positive bias obviously cannot explain the boost in action binding that occurs specifically after non-rewarded outcomes. However, we found that the inter-individual variability of the post-error boost was related to asymmetry of participants’ learning rates. Participants whose learning was more marked for non-rewarded relative to rewarded outcomes also displayed stronger post-error boosts in action binding. While we cannot be sure of the direction of causation underlying this relation, the observed correlation suggests a strong linkage between learning and agency.

Interestingly, this asymmetric effect on sense of agency recalls similar asymmetries in FRN, an EEG component thought to reflect anterior cingulate cortex activity. FRN is stronger after unfavourable outcomes and stronger for participants who tend to learn more from their mistakes (Frank, Woroch, and Curran, 2005). Moreover, similar to our post-error boost in action binding, the FRN was increased only when participants could actually learn, i.e., when they had the opportunity to choose an action that could influence outcomes (Yeung, Holroyd, and Cohen, 2004) or were told that a task was ‘controllable’ compared to ‘uncontrollable’ (Li et al., 2011). These parallels point to a possible link between action binding and FRN, which we will investigate in future research.

The structure of the reversal-learning paradigm inevitably carries some confounds when investigating effects of errors. Specifically, errors occur less frequently than successful, rewarded trials. Furthermore, error trials are often associated with the reversal or rule-change event itself. These additional factors could, of course, contribute part of the post-error agency boost we observed. However, we consider that learning from errors remains the more convincing explanation. First, our analyses comparing post-error action boost with frequency of errors found no significant association. Indeed, the numerical sign of the relation was in the opposite direction to the hypothesis described above. We thus found no evidence that post-error boost in action binding is related to non-specific consequences of errors, such as general arousal from 'oddball' events. Second, in our paradigms, the reversal event was never made explicit to the participant and was never entirely predictable. Finally, Experiment 2 found a significant contrast between learning and random conditions, even though actions, outcomes and reversals were equally present in both conditions. Thus, our design clearly links post-error agency boost to the potential for learning about action-outcome relations.

While sense of agency is usually defined as the feeling of controlling one's actions and their consequences (Haggard and Chambon, 2012), few studies have investigated the contribution to sense of agency of action selection processes and of discriminative ability to control outcomes. One previous study suggested that action-outcome relations had no effect on intentional binding (Desantis, Hughes, and Waszak, 2012). Unlike previous studies, our study involved an element of reward-guided decision-making. Experiment 2 showed that discriminative control of outcomes does influence action binding, but only when this element is present, i.e., when people can learn the relation between their actions and possible outcomes. Thus, we suggest that action binding is a useful implicit measure of *goal-directed agency over outcomes*. Binding measures can thus capture a key feature of the sense of agency in the rich sense of everyday life, i.e., the ability to generate one particular external event, rather than another, through one's own motivated, endogenous action.

People normally make actions for a reason. That is, they choose actions to achieve a desired outcome. They then monitor and evaluate whether the action succeeded or failed in achieving the outcome. Thus, one might intuitively expect a link between adaptive behaviour and sense of agency, yet these two traditions in action control have evolved through largely separate research literatures. We show, for the first time, that an implicit measure of sense of agency is sensitive to errors and to reinforcement-learning features. Our data suggest that when people experience unfavourable outcomes, they feel *more* control, not less, in the next trial. This may initially seem counterintuitive, but it is strongly consistent with the view that sense of agency is related to acquiring and maintaining control over external events.

We hypothesize that sense of agency has an important functional role in adaptive behaviour. We speculate that error feedback might transiently boost participants' feeling of agency, because action failures should strongly motivate the requirement to act appropriately on subsequent occasions and also to learn what actions are now appropriate. Sense of agency could be understood in the context of motivation to improve performance on subsequent actions. The human mind houses a specific cognitive/experiential mechanism to ensure that 'If at first you don't succeed, try and try again' (Hickson, 1936). Our study breaks new ground in linking the subjective experience of agency to the cognitive mechanisms of reinforcement learning.

Acknowledgements

All authors contributed to the study concept and design. S.D.C. programmed the experiment. S.D.C. and H.T. tested, collected data, performed the model-free analysis and drafted the manuscript under the supervision of P. Haggard. H.T. performed the model-based analysis. All authors approved the final version of the manuscript for submission. S.D.C. and T.H. contributed equally to this work.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: P.H. was supported by ERC Advanced Grant (HUMVOL, grant agreement 323943). H.T. was supported by an ENS studentship. S.D.C. was supported by a UCL ICN demonstratorship. V.C. was supported by l'Agence Nationale de la Recherche grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-16-CE37-0012-01.

Bibliography

Balleine, Bernard W and Anthony Dickinson (1998). "Goal-directed instrumental action: contingency and incentive learning and their cortical substrates". In: *Neuropharmacology* 37.4, pp. 407–419.

Bandura, Albert (1989). "Human agency in social cognitive theory." In: *American psychologist* 44.9, p. 1175.

Bishop, C (2006). "Pattern Recognition and Machine Learning". In: *Springer, New York*.

Buehner, Marc J and Gruffydd R Humphreys (2009). "Causal binding of actions to their effects". In: *Psychological Science* 20.10, pp. 1221–1228.

Charles, Lucie, Jean-Rémi King, and Stanislas Dehaene (2014). "Decoding the dynamics of action, intention, and error detection for conscious and subliminal stimuli". In: *Journal of neuroscience* 34.4, pp. 1158–1170.

Cools, Roshan et al. (2002). "Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging". In: *Journal of Neuroscience* 22.11, pp. 4563–4567.

Cravo, Andre M, Peter ME Claessens, and Marcus VC Baldo (2009). "Voluntary action and causality in temporal binding". In: *Experimental brain research* 199.1, pp. 95–99.

– (2011). "The relation between action, predictability and temporal contiguity in temporal binding". In: *Acta Psychologica* 136.1, pp. 157–166.

Danielmeier, Claudia and Markus Ullsperger (2011). "Post-error adjustments". In: *Frontiers in psychology* 2.

Daw, Nathaniel D (2011). "Trial-by-trial data analysis using computational models". In: *Decision making, affect, and learning: Attention and performance XXIII* 23, pp. 3–38.

Dayan, Peter and Yael Niv (2008). "Reinforcement learning: the good, the bad and the ugly". In: *Current opinion in neurobiology* 18.2, pp. 185–196.

Desantis, Andrea, Gethin Hughes, and Florian Waszak (2012). "Intentional binding is driven by the mere presence of an action and not by motor prediction". In: *PLoS One* 7.1, e29557.

Dutilh, Gilles et al. (2012). "Testing theories of post-error slowing". In: *Attention, Perception, & Psychophysics* 74.2, pp. 454–465.

Engbert, Kai and Andreas Wohlschläger (2007). "Intentions and expectations in temporal binding". In: *Consciousness and cognition* 16.2, pp. 255–264.

Frank, Michael J, Brion S Woroch, and Tim Curran (2005). "Error-related negativity predicts reinforcement learning and conflict biases". In: *Neuron* 47.4, pp. 495–501.

Haggard, Patrick (2017). "Sense of agency in the human brain". In: *Nature Reviews Neuroscience* 18.4, pp. 196–207.

Haggard, Patrick and Valerian Chambon (2012). "Sense of agency". In: *Current Biology* 22.10, R390–R392.

Haggard, Patrick, Sam Clark, and Jeri Kalogeras (2002). "Voluntary action and conscious awareness". In: *Nature neuroscience* 5.4, pp. 382–385.

Hickson, William Edward (1936). *The Singing Master*. London, England: Taylor Walton.

Lefebvre, Germain et al. (2016). "Asymmetric reinforcement learning: computational and neural bases of positive life orientation". In: *bioRxiv*, p. 038778.

Li, Peng et al. (2011). "Responsibility modulates neural mechanisms of outcome processing: an ERP study". In: *Psychophysiology* 48.8, pp. 1129–1133.

Metcalfe, Janet and Matthew Jason Greene (2007). "Metacognition of agency." In: *Journal of Experimental Psychology: General* 136.2, p. 184.

Moore, James and Patrick Haggard (2008). "Awareness of action: Inference and prediction". In: *Consciousness and cognition* 17.1, pp. 136–144.

Moore, James W and Sukhvinder S Obhi (2012). "Intentional binding and the sense of agency: a review". In: *Consciousness and cognition* 21.1, pp. 546–561.

Niv, Yael et al. (2012). "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain". In: *Journal of Neuroscience* 32.2, pp. 551–562.

O'Doherty, John P (2014). "The problem with value". In: *Neuroscience & Biobehavioral Reviews* 43, pp. 259–268.

Rangel, Antonio, Colin Camerer, and P Read Montague (2008). "A framework for studying the neurobiology of value-based decision making". In: *Nature Reviews Neuroscience* 9.7, pp. 545–556.

Rolls, Edmund T (2000). "The orbitofrontal cortex and reward". In: *Cerebral cortex* 10.3, pp. 284–294.

Sutton, Richard S and Andrew G Barto (1998). *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge.

Takahata, Keisuke et al. (2012). "It's not my fault: postdictive modulation of intentional binding by monetary gains and losses". In: *PloS one* 7.12, e53421.

Walsh, Eamonn and Patrick Haggard (2013). "Action, prediction, and temporal awareness". In: *Acta psychologica* 142.2, pp. 220–229.

Yeung, Nick, Clay B Holroyd, and Jonathan D Cohen (2004). "ERP correlates of feedback and reward processing in the presence and absence of response choice". In: *Cerebral cortex* 15.5, pp. 535–544.

Yoshie, Michiko and Patrick Haggard (2013). "Negative emotional outcomes attenuate sense of agency over voluntary actions". In: *Current Biology* 23.20, pp. 2028–2032.

5.3 Additional experiment

We have also analyzed the results of another experiment conducted by Steven di Costa. In this experiment, we have varied the action-outcome contingencies. Its results replicated our published article.

5.3.1 Methods

16 participants were tested on the same general procedure as in Experiment 1. The only difference was in the action-outcome contingencies: in some blocks, one key delivered rewarded high-tone with a probability of 0.7 and the other key with a probability of 0.3, while in the other blocks the probabilities were 0.9 and 0.1. Therefore there were four agency conditions: one action binding condition and one outcome binding condition for the 70%/30% contingencies, and again one action binding condition and one outcome binding condition for the 90%/10% contingencies.

Experimental trials were categorized according to three design factors: 1. condition (90%/10% vs. 70%/30%), 2. whether outcome on a given trial was positive or negative. 3. whether outcome on the previous trial was positive or negative. Action binding data and outcome binding data were then subjected to a 2x2x2 ANOVA. To further understand participants' strategy, we used the same computational model as the one described in our article.

5.3.2 Results

In both the 70%/30% and the 90%/10% condition, participants were able to learn the correct action. Similarly to Experiments 1 and 2, the trial number after reversal had a significant effect on participants' proportion of correct choice in both the 70%/30% (F(4, 56) = 48.1, p < 0.0001) and in the 90%/10% (F(4, 56) = 139.7, p < 0.0001) conditions.

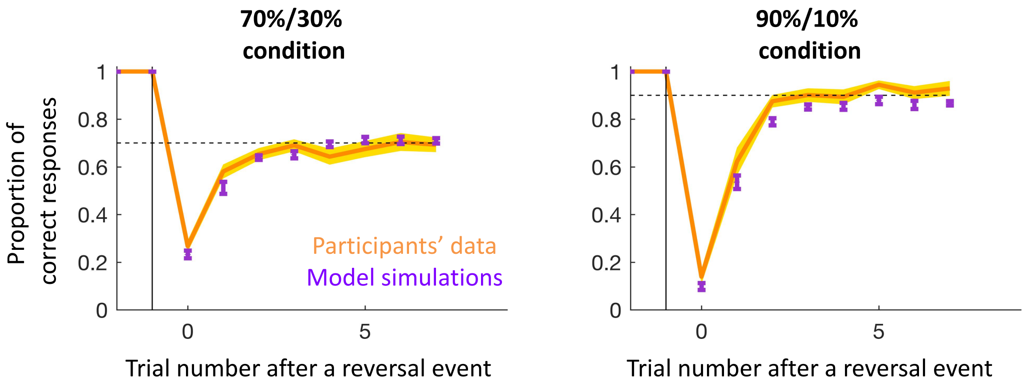


Figure 5.1: Proportion of correct responses before and after a reversal event for the 70%/30% and the 90%/10% conditions. Participants' data are in orange, and predictions of the reinforcement-learning model are in purple. Error-bars represent the standard error of mean.

Consistently with previous findings (Vulkan, 2000), participants followed the probability matching law, i.e. their proportion of correct choices after learning were matched to the probability of reward of the correct action. Indeed in the 70%/30%

condition, participants reached a plateau of 70% of performance a few trials after the reversal while in the 90%/10% condition, the plateau was at 90%.

A 2×2×2 ANOVA on action binding yielded no significant main effect of condition, current outcome or previous outcome, nor any significant interaction (all $p >$ 0.25). The same ANOVA on outcome binding revealed similar results (all $p >$ 0.25). But because this experiment was a replication, we then directly used paired t-tests on the pooled 70%/30% and 90%/10% conditions, to investigate the effect of the previous and current outcomes on action and outcome binding respectively.

As in Experiments 1 and 2, the previous outcome valence had a significant effect on action binding (t_{15} = 2.3, p = .038), with stronger action binding following a negative outcome than following a positive outcome. Regarding outcome binding that was tested only in Experiment 1, the current outcome valence had again a significant effect on outcome binding (t_{15} = -2.4, p = .030), with negative outcomes being more strongly bound towards actions than positive outcomes.

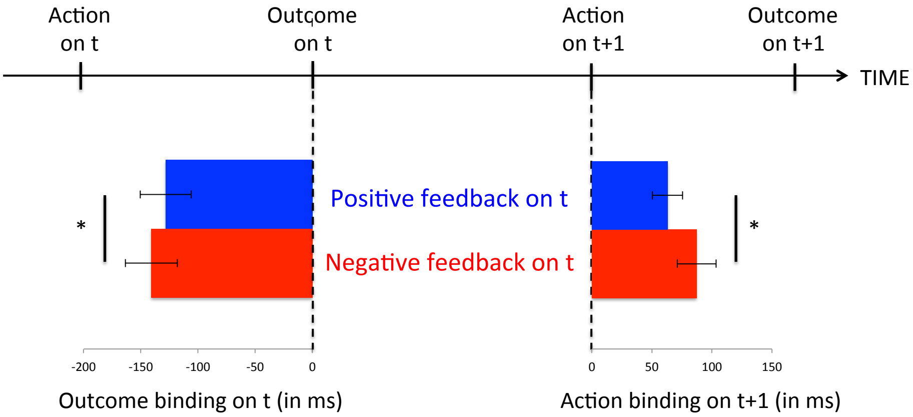


Figure 5.2: On the left, the mean outcome binding (ms) measured for rewarded (blue) and non-rewarded (red) outcomes. On the right, mean action binding (ms) following a rewarded or non-rewarded outcome on the previous trial. Error-bars represent the standard error of mean.

From the fitted parameters, we simulated the model's choices, and we again found a generally good match with participants' performances (see the previous performance figure). Similarly to Lefebvre et al. (2016), Palminteri et al. (2017) and our published article, we also found a higher learning rate for positive outcomes than negative outcomes (paired t-test, $t_{15} = 6.94$, $p < 0.001$). We will develop extensively this learning rate assymetry results and interpret it in Study III.

We finally explored the correlation between the difference in action binding and the learning rate asymmetry. We found a positive correlation between the post-error boost of agency and the normalized learning rate asymmetry, although it was not significant ($R = 0.38$, $p = 0.15$).

In this additional experiment, we have used novel action-outcome contingencies. We replicated that, in a reinforcement-learning environment, negative outcomes led to increased outcome binding, while also increasing action binding in the following trial. Unfortunately, we lacked statistical power to further analyze the difference between the 70%/30% and 90%/10% conditions. We also found a positive, although non significant, correlation between the post-error boost of action binding and the learning rate asymmetry. This experiment generalized our findings to new action-outcome contingencies, although it remained inconclusive about the specific effects of the manipulated contingencies.

5.4 Additional analyses

After the article publication, we further analyzed Experiment 2's results. Experiment 2 consisted of two main conditions: a learning condition in which one key delivered rewarding high tones with a probability of 0.8 and the other key with a probability of 0.2, and a random condition in which the probability of rewarding tones was the same for the two keys (0.5). These conditions were explicitly explained to participants: in the learning condition they were instructed to "find the good key, maximizing the number of high tones", whereas in the random condition they were told, "whichever action is chosen, it will have no influence on the following tone". Furthermore, participants could earn a performance bonus only in the learning condition, thus they had no motivation to follow action-outcome contingencies in the random condition. This experiment thus gives us a unique opportunity to study the impact of an explicit lack of instrumental control on participants' behavior.

5.4.1 Behavioral results

Adaptive behavior is often described as a tendency to switch response when the last action led to a negative outcome and to keeping pressing the same key when it previously led to a positive outcome. We expected participants to adapt their responses accordingly to the previous outcome only in the learning condition, and not in the random condition.

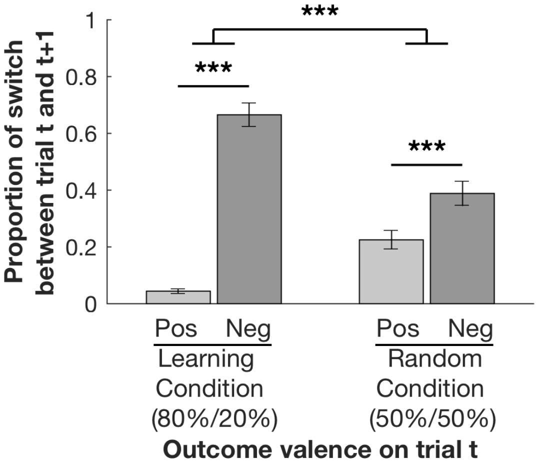


Figure 5.3: The proportion of switch between trial t and t+1 as a function of the valence (positive or negative) of the outcomes seen on the trial t, sorted between the trials in the learning and random conditions. The standard errors appearing on this graph (and the following ones) were calculated across participants, while the stars indicate significance on paired t-tests (*: p < .05; **: p < .01; ***: p < .001).

We computed the proportion of key switch between trial t and t+1, as a function of the outcome valence (positive or negative) on trial t, and as a function of the experimental condition (learning or random condition). We subjected these proportions to a 2×2 ANOVA. We found a highly significant effect of outcome valence ($F_{1,29} = 25.9, p = 1.5 \times 10^{-6}$), with more switching behavior following negative outcomes than following positive outcomes. There was no main effect of experimental condition ($F_{1,29} = 0.36, p = .55$). Crucially there was a significant valence × condition interaction ($F_{1,29} = 9.4, p = 2.7 \times 10^{-3}$).

We performed simple-effect t-tests to further investigate this interaction. In both the learning and the random conditions, non-rewarded outcomes significantly increased the proportion of switch on the following trial compared to rewarded outcomes, but the difference was stronger in the learning condition than in the random condition (learning condition: $t_{29} = 14.3, p = 1.1 \times 10^{-14}$; random condition: $t_{29} = 4.7, p = 5.5 \times 10^{-5}$; difference between conditions: $t_{29} = 8.7, p = 1.1 \times 10^{-9}$).

Although participants were explicitly told their actions had no influence on outcomes in the random condition, they still displayed adaptive behavior, instead of randomly choosing between any key or always choosing the same. But this adaptation was less pronounced than in the learning condition, showing that the instructions did regulate participants' strategy.

5.4.2 Computational results

We then used a reinforcement-learning model to further understand the different strategies implemented by participants in the two conditions, and to investigate whether an explicit lack of control had an impact on the asymmetry between the positive and the negative learning rates.

We applied the same modeling procedure as the one described in the above article, baring a few differences. First, we used this model to fit both conditions in Experiment 2 (and not only the learning condition). Second we fitted four learning rates, instead of two: we used a pair of alphas (α^+ and α^-) to fit the learning condition, and another pair to fit the random condition.

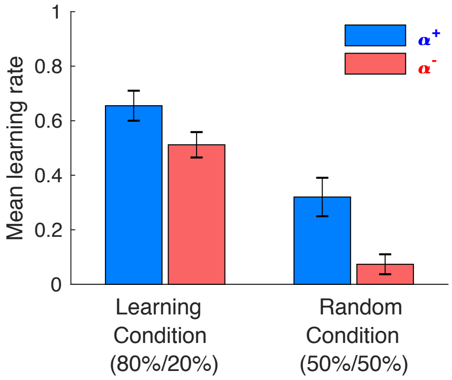


Figure 5.4: Mean learning rates for positive (α^+, in blue) and negative (α^-, in red) outcomes in the learning and random conditions.

We then subjected the learning rates to the same 2×2 ANOVA as the proportions of switch, i.e., the outcome valence and experimental condition as the predictors. We found a significant effect of outcome valence ($F_{1,29} = 8.6, p = 4.0 \times 10^{-3}$), with higher learning rates for positive outcomes than for negative ones. There was a highly significant effect of experimental condition ($F_{1,29} = 17.1, p = 6.8 \times 10^{-5}$), with higher learning rates in the learning than in the random condition. Interestingly we found no valence × condition interaction ($F_{1,29} = 0.02, p = 0.88$).

The asymmetry between α^+ and α^- was thus considered similar between the learning and the random conditions. This result is striking because we will see in the next study that being forced to choose between two options abolished the asymmetry between α^+ and α^-. Therefore a lack of agency in action choice made the learning rate asymmetry vanish, while here an explicit lack of agency over action outcomes did not prevent positive outcomes to be more integrated than negative outcomes.

Still we found learning rates to be higher in the learning than in the random condition. We can interpret it as participants choosing more randomly between the two keys, regardless of previously observed outcomes, in the random than the learning condition. But a change in learning rates can be interpreted differently, as reflecting the environment stability (Behrens et al., 2007). Our reinforcement-learning model allowed us to investigate how far participants' choice was from randomness, as the model computed the probabilities to choose each key on each trial. We could thus compute how likely the participants' choices were to occur, according to the model. We found that the probability of the model selecting the same action as the participant was 0.82 ± 0.02 (mean ± standard error mean) in the learning condition and 0.65 ± 0.03 in the random condition. This difference in predictive power was significant (paired t-test: $t_{29} = 4.8, p = 4.8 \times 10^{-5}$). Participants' choices were thus predicted to a lesser extent by a reinforcement-learning model in the random than in the learning condition, showing that participants' behavior was less outcome-driven and more stochastic in the random condition.

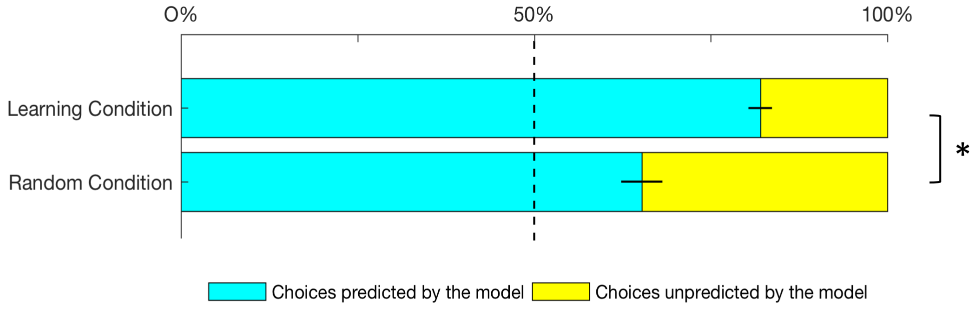


Figure 5.5: Mean predictive power, i.e. percentage of choices predicted by the model, for the learning and the random conditions.

When participants were told they had no control over outcomes, we found that their adaptive behavior was reduced and that a reinforcement-learning model was less able to explain participants' choices, thus making their choices more stochastic, than when participants are instructed to learn action-outcome contingencies. Consistently with Lefebvre et al. (2016)'s findings, we also found a lack of control to have no effect on the learning rate asymmetry.

Chapter 6

Study III

6.1 Introduction

In the previous study, we found participants' learning rates to be higher for positive than negative outcomes for factual learning, i.e., learning from obtained outcomes. Palminteri et al. (2017) found the same result, but also the opposite valence-induced bias for counterfactual learning, as negative counterfactual outcomes were preferentially integrated, relative to positive ones. These results can be generally seen as a choice-confirmation bias.

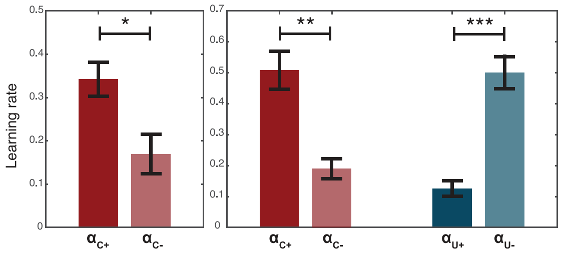


Figure 6.1: The choice-confirmation bias recently found in learning rates. Factual and counterfactual learning rates are respectively denoted α_C and α_{CF}. (Figure reproduced from Palminteri et al., 2017.)

Choice-confirmation is a self-centered bias: we want our choice to be correct, and thus interpret the given outcomes in this light. In our previous study, when we fitted the learning rates independently in Experiment 2's learning and random conditions, we found the same difference between positive and negative learning rates for explicit presence and lack of outcome control. The choice-confirmation bias appeared to persist in a situation in which participants were told that their actions could not control outcomes and that there were no 'correct' or 'incorrect' choice to make.

In this study, we hypothesized the learning rate asymmetry to disappear when participants were forced to select an action, and we investigated the optimality of differential learning rates in various experimental conditions.

6.2 Our draft in preparation

Confirmation bias in instrumental but not observational learning

Héloïse Théro[1], Henri Vandendriessche[1], Patrick Haggard[2], Valérian Chambon[3]* and Stefano Palminteri[1]*

* co-senior authors

[1] *Laboratoire de Neurosciences Cognitives Computationnelles, INSERM-ENS, Département d'Etudes Cognitives, PSL University, Paris, France*

[2] *Institute of Cognitive Neuroscience, University College London, London, UK*

[3] *Institut Jean Nicod, ENS-EHESS-CNRS, Département d'Etudes Cognitives, PSL University, Paris, France*

Corresponding authors: Héloïse Théro (thero.heloise@gmail.com) and Stefano Palminteri (stefano.palminteri@ens.fr)
Laboratoire de Neurosciences Cognitives Computationnelles, École normale supérieure, 29 rue d'Ulm, 75005 Paris, France.

June 30, 2018

Recent findings show that even at the 'low-level' reinforcement learning process, individuals display a choice-confirmation bias, i.e. they preferentially take into account information that confirms their current decision. Beyond classical learning, individuals are also able to learn from merely observing outcomes generated by an external source. We wondered if participants would still display a choice-confirmation bias in the case of observational learning. We analysed two experiments in which the participants' choice was either 'free' or 'forced', and used a computational model adapted to test if outcome valence influences learning. We found that the confirmation pattern previously described can only be found in free-choice trials, as forced-choice trials triggered a valence-impartial learning. A model comparison analysis confirmed this findings, as the winning model had valence-independent learning rates in forced-choice trials.

Keywords

Reinforcement learning; confirmation bias; observational learning; outcome valence; free vs forced-choice; counterfactual outcome.

Introduction

Humans should be able to take impartially into account different information, regardless of its irrelevant features like valence. Classical theories of reinforcement learning assume that action values are learnt via the calculation of a reward prediction error, i.e., the difference between the obtained and the expected outcome, and they suppose that subjects learn similarly, independently of the valence (positive or negative) of the prediction error (Sutton and Barto, 1998). Theoretical simulations ground this supposition for valence-independent learning: having a bias for positive or negative prediction error has been shown to be a sub-optimal strategy in most situations, although differential learning rates were sometimes advantageous (Cazé and van der Meer, 2013).

Yet, recent findings from various paradigms show that humans display a significant valence-induced bias. It generally goes in the direction of preferentially learning from positive, compared to negative prediction error (Lefebvre et al., 2017). This asymmetry may be interpreted as a general optimism bias in human in-

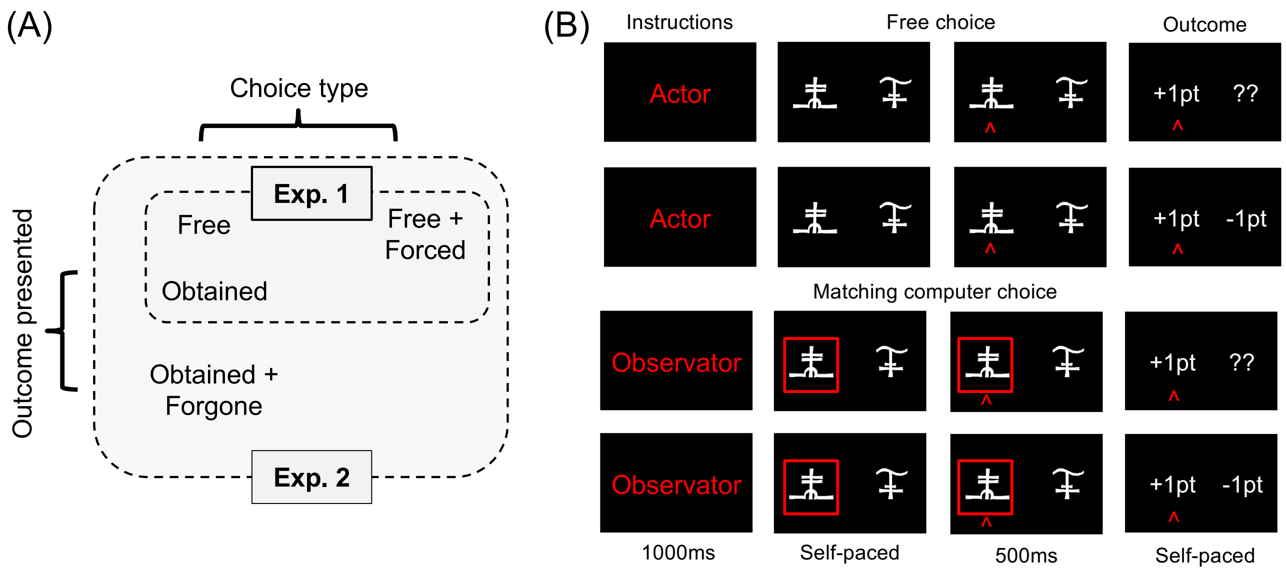


Figure 1: *(A) Experiment 1 was composed of a condition with only free-choice trials, and a condition with intermixed free- and forced-choice trials. Only partial trials were used. In Experiment 2, we had always free- and forced-choice trials intermixed, and there were two conditions: one with partial trials, and one with complete trials.* ***(B)*** *Description of the four trial types implemented in Experiments 1 and 2. In free-choice trials, participants could freely select between the two options, while in forced-choice trials, participants had to match a preselected option. In partial trials, participants were shown only the reward (+1 or -1) associated to the chosen option, while in complete trials, participants were shown the outcome of both the chosen and unchosen option.*

formation integration (Sharot and Garrett, 2016; Kuzmanovic and Rigoux, 2016). A recent article designed an experiment in which participants were shown not only the obtained outcome for their chosen option (factual outcomes), but also the forgone outcome associated with the unchosen option (counterfactual outcomes, Palminteri et al., 2017). For factual outcomes, they have replicated that participants learned preferentially from positive, relative to negative, outcomes. But for counterfactual learning, negative outcomes were preferentially taken into account, relative to positive ones. These results are therefore best explained by a choice-confirmation bias (i.e. people integrate preferentially information that confirms their choice) than a positivity bias.

If participants are more prone to integrate choice-confirming outcome, one can wonder what happens when participants see information without any choice to make. Indeed, individuals are able to learn not only from their own actions and outcomes but also from those that are observed, this observational learning appeared to rely on similar neural mechanisms as classical learning (Burke et al., 2010; Cooper et al., 2012; Burke et al., 2016; Monfardini et al., 2013). It should be noted that these experiments involved learning from observing another person's actions and outcomes.

We conducted two simple instrumental learning tasks in which the participants' choice was either 'free' or 'forced' (Figure 1). We followed a classical operationalization of choice freedom: in an instructed or forced choice, actions are fully specified by external stimuli while a free action occur in underdetermined external environments (Filevich et al., 2013). In a first experiment (N = 24), participants were shown only the factual outcome corresponding to their choices. We hypothesized that in forced-choice, participants would not be subjected to the choice-confirmation bias. In a second experiment (N = 24), the counterfactual outcome, i.e., the outcome associated with the unchosen symbol, was also displayed. Our goal was to replicate our findings, and verify that counterfactual learning would also be not be subjected to the choice-confirmation bias in forced-choice trials.

Results

In the two experiments, participants performed an instrumental learning tasks in which their choices were either 'free' or 'forced'. Participants were instructed to find the symbol associated with a higher probability of reward (i.e., '+1' outcome). In free-choice trials, the participant freely chose between two symbols, whereas in forced-choice trials the computer preselected one symbol and the subject was forced to match this choice. Crucially in forced-choice trials, the two stimuli were pseudo-randomly preselected, thus ensuring equal sampling from both the option associated with high value and the other.

Only the factual outcomes, i.e., the outcomes associated with the chosen stimuli, were shown in Experiment 1 (N = 24). We called these trials 'partial', as only a part of the two possible outcomes was shown (see

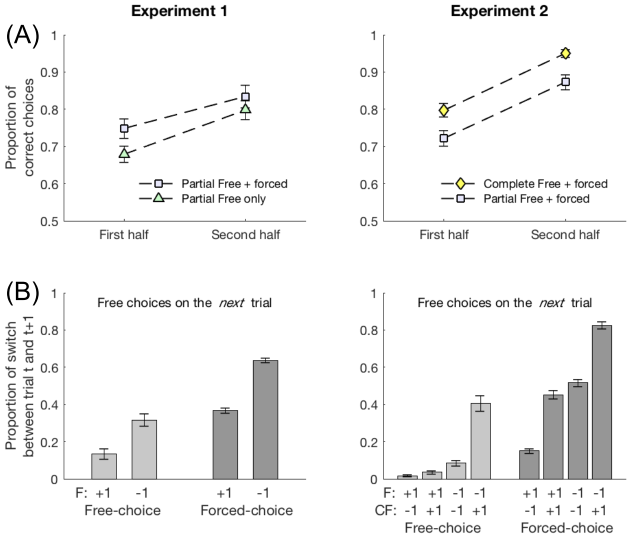


Figure 2: *Behavioral results from Experiments 1 and 2 (respectively shown on the left and right panels).* ***(A)*** *The participants' mean proportion of correct choices for the two conditions, separated between the first and the second halves of the learning block.* ***(B)*** *The proportion of choice switch between trial t and t+1 as a function of the factual (F) and counterfactual (CF) outcomes seen on the trial t, sorted between when the trial t was a free- or a forced-choice trial. For the Experiment 1, this analysis was made on the Partial Free + forced Condition data, as it contained both free- and forced-choice trials. For the Experiment 2, this analysis was made on the Complete Free + forced Condition data, as it contained both factual and counterfactual outcomes. Standard errors shown by the error bars were calculated across participants.*

Figure 1, upper panels). Experiment 1 was composed of 12 blocks, each with a new pair of symbols:

- 6 blocks that each contained 40 free-choice partial trials (the 'Partial Free only' condition)
- 6 blocks that contained 40 free- and 40 forced-choice partial trials, pseudo-randomly intermixed (the 'Partial Free + forced' condition)

In the second experiment (N = 24), we also used trials in which both the factual and the counterfactual outcomes, i.e., the outcomes associated with the chosen and unchosen stimuli, were shown (see Figure 1, lower panels). These trials were called 'complete', as participants had access to the complete information. Experiment 2 was composed of 16 blocks:

- 8 blocks that contained 20 free- and 20 forced-choice partial trials (the 'Partial Free + forced' condition)
- 8 blocks that contained 20 free- and 20 forced-choice complete trials (the 'Complete Free + forced' condition)

Behavior

We first analysed participants' behavior to assess if they were able to perform correctly the task. They were able to find the high-rewarding symbol, as performance (i.e. average correct choice proportion) was significantly higher than chance level in each condition and in each experiment (t-tests against 50%: $t_{23} > 10, p < 10^{-9}$; see Figure 2A).

Performance data was further categorized according to two design factors: whether the trial was part of the first or second half of the learning block; and depending on the condition the trial belonged to. We thus subjected performance to a 2×2 repeated-measure ANOVA. In both experiments, we found a significant effect of the leaning block phase, with performance increasing between the first and second halves of the blocks (Exp. 1: $F_{1,23} = 7.5, p = 7.6 \times 10^{-3}$; Exp. 2: $F_{1,23} = 13, p = 5.3 \times 10^{-4}$; see Figure 2A).

Participants seemed to use information from forced-choice trials and from counterfactual outcomes to improve their performance, as their performances were higher in the 'Partial Free + forced' condition than in the 'Partial Free only' condition, and were also higher in the 'Complete Free + forced' condition than in the 'Partial Free + forced' condition. But these differences were not significant, as we found no main effect of condition (Exp. 1: $F_{1,23} = 1.5, p = 0.22$; Exp. 2: $F_{1,23} = 3.5, p = 0.063$), nor a phase-by-condition interaction effect (Exp. 1: $F_{1,23} = 0.11, p = 0.74$; Exp. 2: $F_{1,23} = 0.018, p = 0.89$).

We can show that participants did pay attention to the outcome associated with the preselected stimuli in forced-choice trials, and with the counterfactual outcomes. The hallmark of adaptive behavior is to switch choice after a negative outcome and to repeat a choice after a positive outcome. We therefore analysed the proportion of choice switch depending on: whether the previous factual outcome was positive or negative; whether the previous trial was a free- or forced-choice; and for Experiment 2 only, whether the counterfactual outcome was positive or negative.

The ANOVAs revealed a main effect of the factual outcome (Exp. 1: $F_{1,23} = 16, p = 1.1 \times 10^{-4}$; Exp. 2: $F_{1,23} = 74, p = 4.2 \times 10^{-15}$). Thus after observing a negative factual outcome in a previous trial, participants indeed switched more often options than after observing a positive factual outcome. Crucially this phenomenon occurred similarly after a free- and a forced-choice trial. Moreover we found a main effect of counterfactual outcome in Experiment 2 ($F_{1,23} = 39, p = 3.6 \times 10^{-9}$), with participants switching more often when the outcome associated with the unchosen option was positive, relative to negative (see Figure 2B).

It should be noted that we found a main effect of choice type ($F_{1,23} > 36, p < 10^{-7}$), with participants switching more after a forced-choice trial than after a free-choice trials. Indeed a symbol was pseudo-

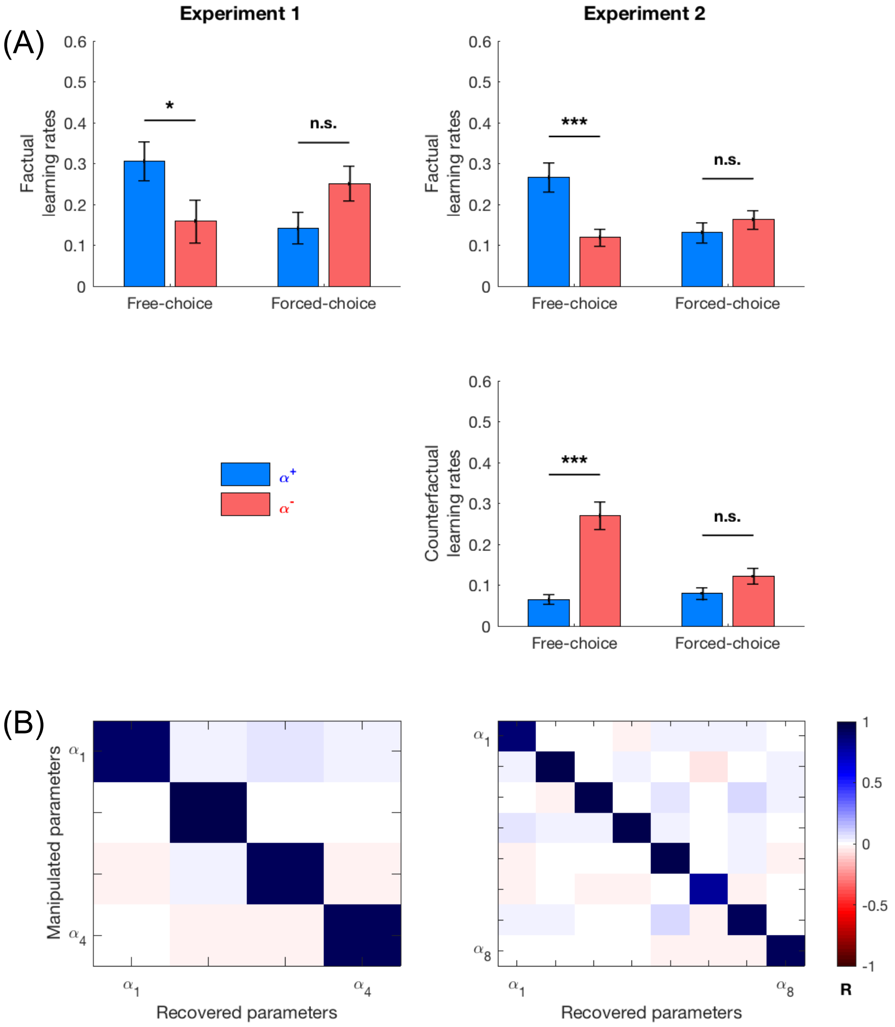


Figure 3: *Parameter results of the 'Full' model from Experiments 1 and 2.* ***(A)*** *We represented the positive learning rates in blue and the negative learning rates in red, fitted from the free- and forced-choice trials. The factual (respectively counterfactual) learning rates are in the upper (respectively lower) panels. Note that Exp. 1 included only factual outcomes, while in Exp. 2, counterfactual outcomes were sometimes displayed, allowing us to fit counterfactual learning rates. The stars indicate significance on paired t-tests (*: p < .05; **: p < .01; ***: p < .001).* ***(B)*** *Parameter recovery results of the 'Full' model with 4 (Experiments 1, left panels) or 8 (Experiment 2, right panels) learning rates. We have represented the matrices of averaged correlation coefficients R of the correlations between the manipulated parameters used to simulate artificial datasets and the recovered parameters from using our fitting procedure on these simulated datasets.*

randomly preselected in forced-choice trials. The 'correct' and the 'incorrect' symbols had thus the same probability to be chosen, causing a high proportion of choice switch on the following trial.

We found that participants efficiently used forced-choice trial and counterfactual information to learn action-outcome contingencies, justifying a computational model learning from both free- and forced-choice trials, and from factual and counterfactual outcomes.

Model parameter analyses

To assess a difference in learning between free- and forced-choice trials, we fitted an established model of reinforcement-learning, with different pairs of positive and negative learning rates (α^+ and α^-) in free- and forced-choice trials, and for factual and counterfactual outcomes in Experiment 2. The resulting learning rates were subjected to the same ANOVAs as the previous proportions of choice switch.

The learning rate asymmetry was different between the free- and forced-choice trials. Indeed we found a significant choice × valence interaction in the Experiment 1 ($F_{1,23} = 7.4, p = 7.6 \times 10^{-3}$), and a significant choice × valence × factuality interaction in the Experiment 2 ($F_{1,23} = 6.8, p = 1.0 \times 10^{-2}$). We performed post-hoc t-tests to further investigate these interactions. The difference between positive and negative learning rates was always significant in free-choice trials (Exp. 1: $t_{23} = 2.5, p = 2.0 \times 10^{-2}$; Exp. 2, factual: $t_{23} = 4.1, p = 4.3 \times 10^{-4}$; and counterfactual: $t_{23} = -6.2, p = 2.6 \times 10^{-6}$), and non-significant in forced-choice trials (Exp. 1: $t_{23} = -2.0, p = 0.055$; Exp. 2, factual: $t_{23} = -1.3, p = 0.20$; and counterfactual: $t_{23} = -1.5, p = 0.14$, see Figure 3A). Although the difference between the positive and negative learning rates was close to significance in Experiment 1, this difference almost completely vanished in Experiment 2.

The ANOVA also revealed a significant valence × factuality interaction in Experiment 2 ($F_{1,23} = 11, p = 1.2 \times 10^{-3}$), replicating Palminteri et al. (2017)'s results. Indeed we can see a confirmatory bias in the free-choice learning rates: participants learned preferentially from positive, relative to negative, factual outcome whereas the opposite pattern appeared for counterfactual outcomes. All other main or interaction effects were non-significant ($p > 0.05$, see Tables 1 and 2).

Table 1: *The F- and p-values from the 2×2 ANOVA on the learning rates fitted on participants' choice from Experiment 1.*

	F-values	p-values
Choice	.17	.68
Valence	.015	.90
Choice×Valence	7.4	.0076

Table 2: *The F- and p-values from the 2×2×2 ANOVA on the learning rates fitted on participants' choice from Experiment 2.*

	F-values	p-values
Choice	3.5	.064
Valence	.59	.44
Factuality	1.8	.18
Choice×Valence	.88	.35
Choice×Factuality	.77	.38
Valence×Factuality	11	.0012
Choice×Valence×Factuality	6.8	.010

We then used a parameter recovery procedure to assess whether these results were parameter fitting artefacts. We applied the same parameter fitting procedure to simulated datasets and found that on average parameters were significantly well recovered (all $R_s > 0.78$, all $p_s < 10^{-3}$). Crucially, our fitting procedure introduced no spurious correlations between the manipulated parameters and the other recovered parameters (all $-0.058 < R_s < 0.082$, all $p_s > 0.43$, see Figure 3B).

Parcimony-driven parameter reduction

Although we found no valence-driven difference in forced-choice learning rates, it is possible that participants had opposite biases that cancelled one another on aggregate measures. We therefore ran a parcimony-driven parameter reduction to see if fitting different learning rates in forced-choice was important to predict participants' data. The 'Full' model (i.e., the model with 4 α in Exp. 1 and 8 α in Exp. 2, whose parameters are shown in Figure 3A) corresponds to the one whose parameters were described previously. In the 'Intermediate' model, the negative and positive learning rates are set to be equal on forced-choice trials only. Finally, in the 'Reduced' model, the negative and positive learning rates are always equal on free- and forced-choice trials (see Figure 4A)..

We first compared the models using a Bayesian model selection (Daunizeau, Adam, and Rigoux, 2014) based on the Bayesian Information Criterion (BIC). The 'Intermediate' model was found to better account for the data compared to other models, as its average posterior probability was higher than the posterior probabilities of the other models. Moreover, its exceedance probability, i.e., the probability of this model being more likely than any other model, was 0.81 in Experiment 1 and 1.0 in Experiment 2 (see Figure 4B).

The BIC tend to favor overly simple models because it relies on specific assumptions (Bishop, 2006; Daw, 2011). We thus also used a cross-validation procedure. We can see that the 'Reduced' model under-fitted the data as its cross-validation likelihood was significantly lower than the 'Intermediate' model's (Experiment 1: $t_{23} = -3.9, p = 6.7 \times 10^{-4}$; Experiment 2: $t_{23} = -4.0, p = 5.1 \times 10^{-4}$). In Experiment 1, the difference between the 'Intermediate' (3 α) and 'Full' (4 α) models was not significant ($t_{23} = 1.4, p = 0.18$), although a parsimony approach would recommend to keep the simpler model at equal performance. Still Experiment 2's results indicated that the 'Full' (8α) model over-fitted participants' choices, as its cross-validation likelihood was lower than the 'Intermediate' (6α) model's ($t_{23} = -3.7, p = 1.1 \times 10^{-3}$; see Figure 4C).

Table 3: *The mean and standard errors of the winning model parameters for Experiments 1 and 2.*

		Factual			Counterfactual		
	β	α^{+}_{free}	α^{-}_{free}	α_{forced}	α^{+}_{free}	α^{-}_{free}	α_{forced}
Exp. 1	4.2	.35	.14	.13	-	-	-
	(±0.50)	(±.063)	(±.054)	(±.036)			
Exp. 2	6.3	.30	.11	.14	.065	.27	.089
	(±0.60)	(±.035)	(±.020)	(±.022)	(±.011)	(±.033)	(±.011)

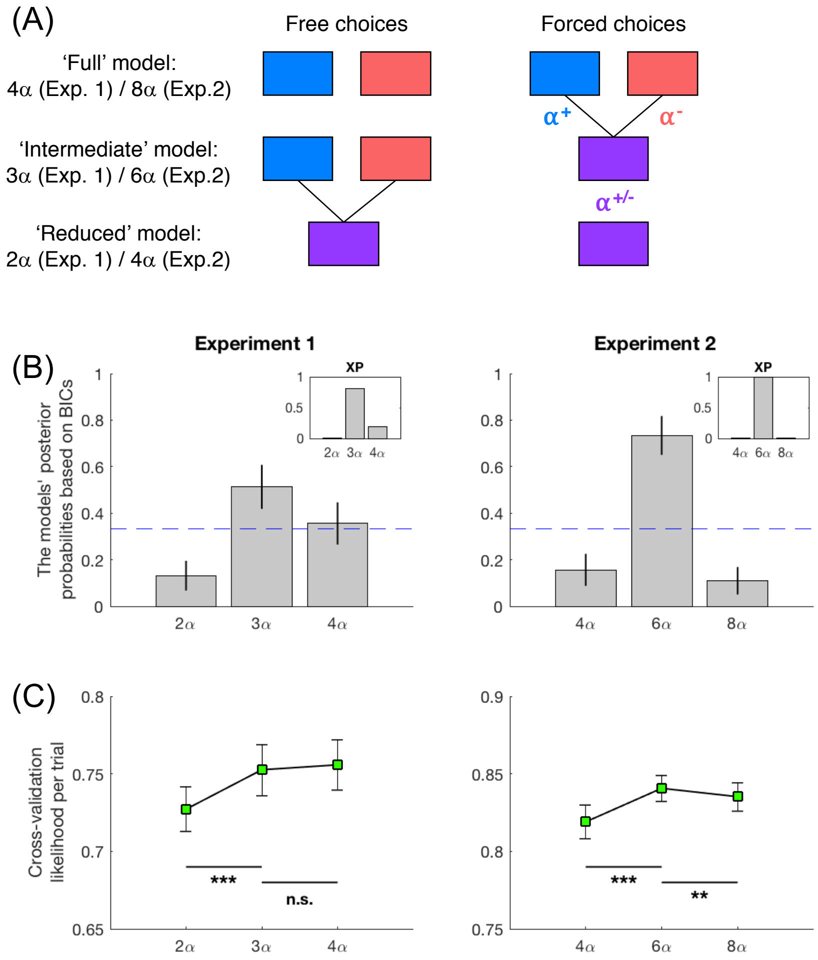


Figure 4: *Model comparison results from Experiments 1 and 2.* ***(A)*** *illustration of the model space. The 'Full' model had different positive and negative learning rates in both free- and forced-choice trials. In the 'Intermediate' model, the positive and negative learning rates were set to be equal in forced-choice trials, while in the 'Reduced' model, the positive and negative learning rates were always set to be equal. The 'Reduced' model is thus nested within the 'Intermediate' model, which is itself nested within the 'Full' model. Note that in Experiment 2, this parameter reduction occurred for both factual and counterfactual learning rates.* ***(B)*** *The expectations and the variances of the posterior probability for each model, based on the Bayesian Information Criterion (BIC) values, with the exceedance probability (XP) for each model in small windows.* ***(C)*** *The average likelihood per trial after applying a cross-validation procedure.*

Our model comparison showed that the 'Intermediate' model is the most likely. The parameters of this winning model are shown in Table 3.

Parameter optimality

Then we simulated models with different learning rate patterns to understand how parameter values can affect performance in our task. We set learning rates to be either choice-confirmatory ('Conf'), valence-neutral ('Neut') or choice-disconfirmatory ('Disc'), and the learning rate patterns could be different in free-choice and forced-choice trials (see Figure 5C and Table 4).

Cazé and van der Meer (2013) have found that different learning rate patterns can be advantageous for certain reward contingencies, that they called 'low-reward' and 'high-reward' (when the reward probabilities are respectively both low or both high for the two possible actions). We have thus used in our experiments low-reward conditions in which the reward contingencies were 0.4 and 0.1, and high-reward conditions in which the contingencies were 0.9 and 0.6. Replicating Cazé and van der Meer (2013)'s findings, we also found the choice-confirmatory models to outperform the other models in low-reward conditions, and the choice-disconfirmatory models to have better performances in the high-reward conditions (see Figure 5).

When we looked at the general performance across both conditions, we found that the highest performing model was the model corresponding to the participants' learning rate patterns, i.e., the 'Conf & Neut' model whose learning rates were choice-confirmatory in free choices and valence-neutral in forced choices (see Figure 5). Even in Experiment 1, the 'Conf & Neut' model had a performance of 85,1%, while the closest one, the

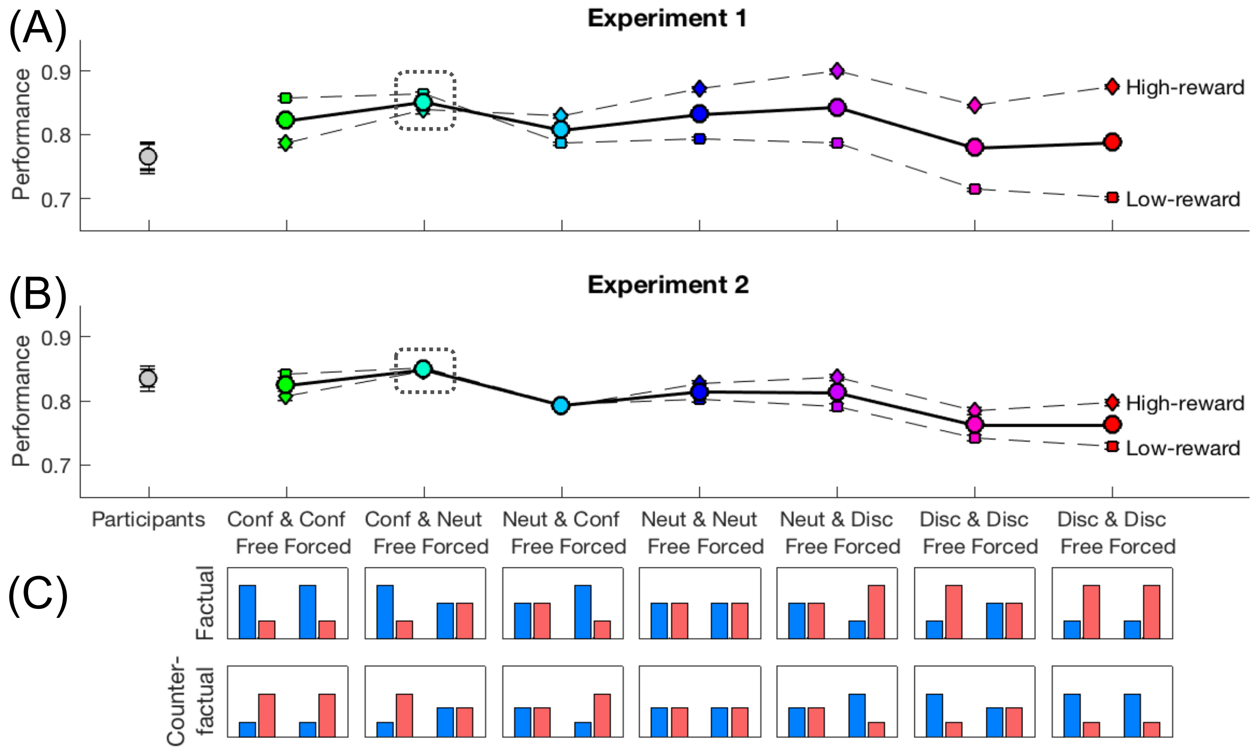


Figure 5: *Parameter optimality was tested by simulating models with different learning rate patterns in Experiments 1 (A) and 2 (B). The models were named depending on their learning rate patterns, shown in (C). For example, 'Conf & Neut' designated a model with choice-confirmatory learning rates in free-choice trials and valence-neutral learning rates in forced-choice trials. The diamonds and the squares correspond to the performance in high- and low-reward conditions respectively. The circles correspond to the performance averaged across the two conditions. The performance corresponding to the participants' pattern of learning rates was highlighted with a dashed square. Participants' actual performances were also shown in grey. Error-bars were plotted, although they were often too small to be seen.*

'Neut & Disc' model had a lower performance of 84,3%. Therefore the learning rate patterns we found in our participants can be said to be optimal.

Interestingly, the performances of the 'Conf & Neut' model were also quite close on the high- and low-reward conditions.. In Experiment 1, the 'Conf & Neut' model had the smaller difference in performance (2%) between the high and low-reward conditions, while this difference was over 4% for the other models. In Experiment 2, the performance difference was 0.46% for the 'Conf & Neut' model. It was only smaller for the 'Neut & Conf' model (0.27%), as this difference was over 2% for the rest of the models. Therefore the 'Conf & Neut' model not only outperformed the others, but also had stable performances in our two different conditions. Crucially we could also see in our participants' data similar performances in high- and low-conditions (paired t-tests for Exp. 1: $t_{23} = 0.25, p = 0.80$; for Exp. 2: $t_{23} = -0.027, p = 0.98$).

Discussion

Two cohorts of healthy adult participants performed an instrumental learning task intermixing free-choice and forced-choice trials. We have investigated the relation between the type of choice and the subsequent reinforcement learning processes. We replicated the choice-supportive bias usually found when participants were free to choose between two alternatives (Lefebvre et al., 2017; Palminteri et al., 2017). Crucially, we found this valence bias to be absent in forced-choice trials. This result was further supported by model comparison analyses.

Another experiment has computationally investigated how learning from free and forced choice outcomes may differ. Cockburn, Collins, and Frank (2014) have simulated a reinforcement learning model similar to ours but they used it to predict participants' post-learning choice rate, whereas we have predicted participants' trial-by-trial choices during learning. They found positive outcomes in free choices to have more impact on learning that negative outcomes. Consistently with our findings, they modelled no difference between positive and negative outcomes in forced choices. This model was able to explain why humans exhibit a preference for freely chosen options.

A vast literature in psychology have investigated how choice can affect preferences and explicit memory. For example, choice-induced preference change, for example, refers to an observation that after choosing between two similarly valued items, participants rate the selected item better than they initially did, and the rejected option as worse (Brehm, 1956). As a control, it was shown that preferences were not altered when a computer instructed the participants decision (Sharot, Velasquez, and Dolan, 2010). Choice was also shown to alter memory (Mather and Johnson, 2000; Mather, Shafir, and Johnson, 2000). Participants were asked to virtually choose between two potential roommates, each with some positive and negative features. When asked after their choice, participants displayed a choice-supportive memory distortion: they tended to attribute, both correctly and incorrectly, more positive features to the chosen person, and more negative features to its competitor (Mather, Shafir, and Johnson, 2000). As the authors said, the consequences of such bias is that 'it is problematic for learning from past experience'. Still, this effect also disappeared when one option was assigned, and not chosen by the participant (Mather, Shafir, and Johnson, 2003). Our results furthermore support the idea that choice-confirmation biases are pervasive in human cognition (Nickerson, 1998).

A few studies have also shown that free choice result in robust enhancements of declarative memory (Voss et al., 2011; Murty, DuBrow, and Davachi, 2015). This behavioral results are consistent with different human neuroimaging studies founding greater BOLD response for free, compared with instructed, actions in areas involved in action planning, as the supplementary motor area (Krieghoff et al., 2009; Filevich et al., 2013) and the anterior cingulate cortex (Lau et al., 2004; van Eimeren et al., 2006). Event-related potential analyses in electroencephalogram showed that choice freedom can also alter processing of action outcomes. The auditory N1 and the feedback-related negativity were enhanced in tasks involving a free, compared with an imposed, choice (Yeung, Holroyd, and Cohen, 2004; Yu and Zhou, 2006; Caspar et al., 2016). In our experiments, we have observed that participants could learn as well from free and forced choices. It should be noted that forced choice outcomes could be used to guide subsequent free choices in our task. This may be the reason why participants could learn as well, although differently, from the two trial types.

Interestingly, a neuroimaging study has found that activity for unexpected vs. expected reward was stronger in the right striatum in active learning. In contrast activity in the hippocampus was bilaterally enhanced in observational learning (Bellebaum et al., 2012). Another neuroimaging experiment has found that anticipating an opportunity for choice was associated with increased activity in a network of brain regions thought to be involved in reward processing (Leotti and Delgado, 2011). A recruitment of different brain structures may explain the difference in learning bias we found between free- and forced-choice trials.

Similarly to Cazé and van der Meer (2013), we found the optimistic models to outperform the pessimistic models in the low-reward condition, while this was reversed in the high-reward condition. We have shown here that the participants' learning rate pattern was not biasing them to suboptimal performances, but rather guaranteeing high and stable performances across conditions. We can thus interpret our participants' parameter values as being optimal in our setting, and not as emerging from a maladaptive cognitive bias. However it should be stressed that we were using unusual contingencies, and a stationary setting (i.e., the symbol values were fixed during the learning blocks), which prevents us from generalizing our findings too far. Indeed, an optimistic model was shown to have worse performance than a rational model in a changing environment, and the participants with a higher optimistic bias were worse than the participants having a lower or nonexistent bias (Palminteri et al., 2017).

In summary, by investigating free- and forced-choice learning, the current experiments demonstrate that participants display a choice-confirmation learning pattern in free choices and are impartial to outcome valence in forced choices. This absence of low level reinforcement-learning bias in forced choices may help understand why taking a third-person perspective comes with benefits.

Methods

Participants

This study included two experiments. In each, we tested N = 24 participants (Experiment 1: 13 males, mean age = 25.1 ± 0.8; Experiment 2: 9 males, mean age = 23.9 ± 0.5). The local ethics committee approved the study. All participants gave written informed consent before inclusion in the study, which was carried out in accordance with the declaration of Helsinki (1964, revised 2013). The inclusion criteria were being older than 18 years, reporting no history of neurological or psychiatric disorders and a normal or corrected-to-normal vision. They were paid 10, 15 or 20 euros, depending on the number of points they had accumulated during the experiment.

General procedure

Participants performed a probabilistic instrumental learning task based on previous studies (Lefebvre et al., 2017; Palminteri et al., 2017). Briefly, the task involved choosing between two cues that were associated with stationary reward probability. The possible outcomes were either winning or losing one point. Participants were encouraged to accumulate as many points as possible and were informed that one cue would result in winning more often than the other. They were given no explicit information about the exact reward probabilities, which they had to learn from trials and errors.

Participants were informed that some trials (indicated by the word 'observateur', i.e., 'observer' in French) would be observational trials, meaning that the observed outcome would not be accumulated to their number of points but would allow them to gain knowledge on what would have happened if they had chosen this cue. In Experiment 2, participants were also informed that in some blocks, they would see the outcome associated with the unchosen cue, although they would only accumulate the points of the chosen outcome.

Conditions

Four types of trials were used in this study (see Figure 1). In free-choice trials, participants could freely select between the two possible options, while in forced-choice trials, participants had to match a preselected option. In partial trials, participants were shown only the outcome ('+1' or '-1') associated to the chosen option, while in complete trials, participants were shown the outcome of both the chosen and unchosen options. Experiment 1 was composed of two conditions: a condition with only partial free-choice trials (each block of this condition lasted 40 trials) and a condition with intermixed partial free- and forced-choice trials (each block lasted 40 + 40 = 80 trials). Experiment 2 first parallels the condition with intermixed partial free- and forced-choice trials, and added a condition with intermixed complete free- and forced-choice trials. For the sake of duration, we decided to half the number of trials in Experiment 2 (20 free-choice trials + 20 forced-choice trials in each block).

Blocks

A new pair of cues was used at the start of every block, and participant had to learn from scratch the correct cue over the course of a number of trials. The free- and forced-choice trials experienced by the participant were pseudo-randomly intermixed within a block, and the same pair of cues was used in the free- and forced-choice trials.

In Experiment 1, participants underwent twelve blocks of $\approx$ 3 min (each condition included 6 blocks). Independently from the condition, six blocks were 'high-reward' and six were 'low-reward'. During high-reward blocks, one of the two cues presented produced a gain (+1 point) with probability 0.9 and a loss (-1 point) with a probability of 0.1. The other cue was associated with a probability of 0.6 to produce a gain, and thus a loss probability of 0.4. Regardless of participants' actions, the high-reward blocks led to accumulate more gains than losses. During low-reward blocks, one cue was associated with a gain probability of 0.4, and the other one with a gain probability of 0.1, leading to a majority of losses over gains. In Experiment 2, each condition was composed of eight blocks of $\approx$ 2 min. Half of them were high-reward blocks and the other half were low-reward blocks. The low- and high-reward blocks were associated with the same contingencies as in Experiment 1.

The first block was preceded by a short training (60 trials for Experiment 1; 40 trials for Experiment 2). To ensure participants would not be biased to expect more positive or negative outcomes in the experiment, the action-outcome contingencies during the training block were 0.5 for both possible cue choices.

Trial structure

Trials began by a fixation cross, except when free- and forced-choice were intermixed, in which case the word 'acteur' (respectively 'observateur') appeared for 500ms to indicate the beginning of a free-choice (resp. forced-choice) trial (see Figure 1). A pair of two cues was then presented. The side (right or left) on which the cues appeared was pseudo-randomly chosen on each trial. Participants made their choice by pressing the right or left arrow with their right hand. When the trial was forced-choice, the preselected cue was surrounded by a square. Participants had to press the corresponding arrow in order to move on (nothing happened if they tried to press the other arrow).

Crucially, the cues were preselected to assure equal sampling: in half of the forced-choice trials cue A was preselected, while in the other half cue B was preselected. The obtained outcome was then presented on the side of the chosen cue. When the trial was a complete trial, the foregone outcome was also shown on the side of the unchosen cue. To ensure that participants paid attention to outcomes, even in the case of forced-choice trial, they were asked to press the up arrow when winning a point and the down arrow when losing a point. The choice and the outcome confirmation timing were both self-paced.

The 'full' computational model

We will first describe the 'full' model, whose parameters are shown in Figure 3. According to the reinforcement-learning algorithm, each of the 2 possible stimuli is associated with an internal value called a Q-value (Sutton and Barto, 1998). The Q-values were set as 0 at the beginning of each block, corresponding to the a priori expectation of an equal probability of a positive or negative outcome. Value updating is based on the concept of prediction error, which measures the discrepancy between actual outcome and the expected outcome for the chosen cue, i.e., the chosen Q-value:

$$\delta_{chosen}(t) = O_{factual}(t) - Q_{chosen}(t)$$

where $O_{factual}(t)$ represents the factual outcome on trial t.

The prediction error is then used to update the chosen Q-value:

$$Q_{chosen}(t+1) = Q_{chosen}(t) + \alpha \times \delta_{chosen}(t)$$

where α represents the learning rate parameter.

In the complete condition experienced in Experiment 2, participants were found to learn from both the factual and counterfactual outcomes so in these trials the unchosen Q-value was also updated with the counterfactual outcome using to the same rule:

$$\delta_{unchosen}(t) = O_{counterfactual}(t) - Q_{unchosen}(t)$$
$$Q_{unchosen}(t+1) = Q_{unchosen}(t) + \alpha \times \delta_{unchosen}(t)$$

We set different learning rates, α^+ and α^-, to reflect different updating processes after a positive or negative outcome (Lefebvre et al., 2017; Palminteri et al., 2017). Because we were interested in the specific effect of forced choice on learning, we fitted different pairs of asymmetrical learning rates in free- and forced-choice trials, and for factual and counterfactual outcomes in Experiment 2. The 'full' model thus had 4 α_s in Experiment 1, and 8 α_s in Experiment 2.

In the reinforcement learning framework, the stimuli with the higher Q-value is more likely to be selected. The probability to choose a stimulus will be computed through a softmax function:

$$p_{stimA} = \frac{e^{\beta \times Q_{stimA}}}{e^{\beta \times Q_{stimA}} + e^{\beta \times Q_{stimB}}}$$

where β is the exploitation intensity parameter, which represents the strength of the Q-values' effect on choice selection. We fitted a unique parameter β for all trial and outcome type, to avoid biasing the learning rate comparisons.

The other computational models

We created two simpler versions of the 'Full' model presented above (see Figure 4A):

- an 'Intermediate' model in which the negative and positive learning rates are set to be equal in forced-choice trials, thus leading to 3 α_s in Experiment 1 and 6 α_s in Experiment 2;
- a 'Reduced' model in which the negative and positive learning rates are set to be equal in both free- and forced-choice trials, thus leading to 2 α_s in Experiment 1 and 4 α_s in Experiment 2.

Note that the number of learning rate parameters was always two times higher in Experiment 2 than in Experiment 1, because in Experiment 2 we fitted separately factual and counterfactual learning rates.

Parameter fitting

We fitted the model parameters based on participants' choices on each free-choice trial, independently for each participant. We used a maximum posterior approach (or MAP, Bishop, 2006), to avoid degenerate parameter estimates. The best parameters chosen were therefore those that maximizing the logarithm of the posterior probability (LPP):

$$ln(p(\theta|Choice_{1:N})) \propto ln(p(\theta)) + \sum_{t=1}^{N} ln(p(Choice_t|\theta))$$

where θ represents our parameter set, N is the total number of trials in the experiment, and $p(Choice_t|\theta)$ is the probability that the model would choose the same stimulus as the participant on trial t. To maximize the LPP with respect to θ, we used Matlab's 'fmincon' function with the ranges: $0 < \beta < \infty$ and $0 < \alpha_i < 1$.

The parameter prior probabilities were based on Daw et al. (2011), and we used a gamma distribution with hyperparameters 1.2 and 5 for the β parameter and a beta distribution with hyperparameters 1.1 and 1.1 for the α parameters. To avoid biasing the learning rate comparisons, the same priors were used for all learning rates.

Parameter recovery

We then used a parameter recovery analysis to ensure that our learning rate results were not an artefact from our parameter fitting procedure. Our goal was to assess our capacity of recovering the correct parameters using simulated datasets.

We first simulated performance on our two behavioral tasks using virtual participants in which one learning rate value was being randomly drawn from a uniform distribution between 0 and 1. We then averaged the correlation coefficients R and p-values from 100 correlations between the manipulated parameter and the parameter recovered from the fitting procedure applied to the simulated data set (see Meyniel et al., 2016 for an example of this procedure). This analysis was conducted on all the learning rate parameters of the 'Full' model.

Bayesian Integration Criterion

The logarithm of the parameter posterior probability was used to compute the Bayesian Information Criterion (BIC, Schwarz, 1978) for each model and each participant, as followed:

$$BIC = k \times ln(N) - 2 \times ln(p(\theta_{MAP}|Choice_{1:N}))$$

where k is the number of parameters, and $ln(p(\theta_{MAP}|Choice_{1:N}))$ is the logarithm of the posterior probability (LPP) of the MAP parameters given the participant's choice data.

BIC were then compared between the 'Full', 'Intermediate' and 'Reduced' models to the verify whether the extra learning rate parameters were justified by the data. As an approximation of the model evidence, individual BICs were fed into the MBB-VB toolbox (Daunizeau, Adam, and Rigoux, 2014), a procedure that estimate how likely it is that a specific model generated the data of a randomly chosen subject (the posterior probability of a model) as well as the exceedance probability of one model being more likely than any other model.

Cross-validation

As the BIC do not take account of the uncertainty in the model parameters, it tend to favour overly simple models (Bishop, 2006; Daw, 2011). To assess our true risk of under-fitting our data, we used a cross-validation procedure. In our two experiments, we had different experimental sessions, that were separated by short breaks (3 sessions in Experiment 1, and 4 sessions in Experiment 2).

For each participant and each session, we fitted the model parameters to participants' choices from all other sessions by the same MAP procedure described before. Given these parameters, we then calculated the likelihood of the data in the held-out session. The total likelihood of the data of each participant was then divided by the number of trials in the held-out session to obtain the average choice likelihood per trial.

Parameter optimality

To test the parameter optimality, the models with different learning rates underwent the same experimental conditions as participants did. We simulated the models 1,000 times for each participant (we thus ran $1,000 \times 24$ simulations in each experiment). On each trial, the outcome given to the model was the one associated with the model's choice, and not the participant's. Simulations were used to provide aggregated measures of models' performance.

The learning rate values used for the simulations are described in Table 4 and were chosen to be close to the participants' averaged MAP learning rate values. The exploitation intensity parameter β was set to 5, a value also close to the participants' MAP exploitation intensity parameter.

Table 4: *The learning rate values used to simulate models with a choice-confirmatory ('Conf'), valence-neutral ('Neut') or choice-disconfirmatory ('Disc') pattern on free- or forced-choice trials.*

	Factual		Counterfactual	
	α^+	α^-	α^+	α^-
'Conf'	0.3	0.1	0.1	0.3
'Neut'	0.2	0.2	0.2	0.2
'Disc'	0.1	0.3	0.3	0.1

Statistical analyses

The ANOVAs were performed on R (version 3.3.2) through the 'aov' function. Paired t-tests and correlation tests were performed on Matlab R2017a through the respective functions 'ttest' and 'corrcoef'.

Acknowledgements

All authors contributed to the study concept and design. H.V. programmed the experiment. H.T. tested, collected data, performed all analysis and drafted the manuscript under the supervision of V.C. and S.P.. All authors approved the final version of the manuscript for submission.

Declaration of conflict

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

H.T. is supported by a PSL/ENS PhD scholarship. V.C. is supported by l'Agence Nationale de la Recherche grants ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL, and ANR-16-CE37-0012-01. S.P. was supported by a Marie Sklodowska-Curie Individual European Fellowship (PIEF-GA-2012 Grant 328822) and is currently supported by an ATIP-Avenir grant.

Bibliography

Bellebaum, C et al. (2012). "The neural coding of expected and unexpected monetary performance outcomes: Dissociations between active and observational learning". In: *Behavioural brain research* 227.1, pp. 241–251.

Bishop, C (2006). "Pattern Recognition and Machine Learning". In: *Springer, New York*.

Brehm, Jack W (1956). "Postdecision changes in the desirability of alternatives." In: *The Journal of Abnormal and Social Psychology* 52.3, p. 384.

Burke, Christopher J et al. (2010). "Neural mechanisms of observational learning". In: *Proceedings of the National Academy of Sciences* 107.32, pp. 14431–14436.

Burke, Christopher J et al. (2016). "Partial adaptation of obtained and observed value signals preserves information about gains and losses". In: *Journal of Neuroscience* 36.39, pp. 10016–10025.

Caspar, Emilie A et al. (2016). "Coercion changes the sense of agency in the human brain". In: *Current biology* 26.5, pp. 585–592.

Cazé, Romain D and Matthijs AA van der Meer (2013). "Adaptive properties of differential learning rates for positive and negative outcomes". In: *Biological cybernetics* 107.6, pp. 711–719.

Cockburn, Jeffrey, Anne GE Collins, and Michael J Frank (2014). "A reinforcement learning mechanism responsible for the valuation of free choice". In: *Neuron* 83.3, pp. 551–557.

Cooper, Jeffrey C et al. (2012). "Human dorsal striatum encodes prediction errors during observational learning of instrumental actions". In: *Journal of cognitive neuroscience* 24.1, pp. 106–118.

Daunizeau, Jean, Vincent Adam, and Lionel Rigoux (2014). "VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data". In: *PLoS computational biology* 10.1, e1003441.

Daw, Nathaniel D (2011). "Trial-by-trial data analysis using computational models". In: *Decision making, affect, and learning: Attention and performance XXIII* 23, pp. 3–38.

Daw, Nathaniel D et al. (2011). "Model-based influences on humans' choices and striatal prediction errors". In: *Neuron* 69.6, pp. 1204–1215.

Filevich, Elisa et al. (2013). "Brain correlates of subjective freedom of choice". In: *Consciousness and Cognition* 22.4, pp. 1271–1284.

Krieghoff, Veronika et al. (2009). "Dissociating what and when of intentional actions". In: *Frontiers in Human Neuroscience* 3, p. 3.

Kuzmanovic, B and L Rigoux (2016). "Optimistic belief updating deviates from Bayesian learning". In: *Available at SSRN: http://ssrn. com/abstract_id* 2810063.

Lau, Hakwan C et al. (2004). "Attention to intention". In: *science* 303.5661, pp. 1208–1210.

Lefebvre, Germain et al. (2017). "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4, p. 0067.

Leotti, Lauren A and Mauricio R Delgado (2011). "The inherent reward of choice". In: *Psychological science* 22.10, pp. 1310–1318.

Mather, Mara and Marcia K Johnson (2000). "Choice-supportive source monitoring: Do our decisions seem better to us as we age?" In: *Psychology and aging* 15.4, p. 596.

Mather, Mara, Eldar Shafir, and Marcia K Johnson (2000). "Misremembrance of options past: Source monitoring and choice". In: *Psychological Science* 11.2, pp. 132–138.

– (2003). "Remembering chosen and assigned options". In: *Memory & Cognition* 31.3, pp. 422–433.

Meyniel, Florent et al. (2016). "A specific role for serotonin in overcoming effort cost". In: *Elife* 5.

Monfardini, Elisabetta et al. (2013). "Vicarious neural processing of outcomes during observational learning". In: *PloS one* 8.9, e73879.

Murty, Vishnu P, Sarah DuBrow, and Lila Davachi (2015). "The simple act of choosing influences declarative memory". In: *Journal of Neuroscience* 35.16, pp. 6255–6264.

Nickerson, Raymond S (1998). "Confirmation bias: A ubiquitous phenomenon in many guises." In: *Review of general psychology* 2.2, p. 175.

Palminteri, Stefano et al. (2017). "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8, e1005684.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *The annals of statistics* 6.2, pp. 461–464.

Sharot, Tali and Neil Garrett (2016). "Forming beliefs: Why valence matters". In: *Trends in cognitive sciences* 20.1, pp. 25–33.

Sharot, Tali, Cristina M Velasquez, and Raymond J Dolan (2010). "Do decisions shape preference? Evidence from blind choice". In: *Psychological science* 21.9, pp. 1231–1235.

Sutton, Richard S and Andrew G Barto (1998). *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge.

van Eimeren, Thilo et al. (2006). "Implementation of visuospatial cues in response selection". In: *Neuroimage* 29.1, pp. 286–294.

Voss, Joel L et al. (2011). "Hippocampal brain-network coordination during volitional exploratory behavior enhances learning". In: *Nature neuroscience* 14.1, p. 115.

Yeung, Nick, Clay B Holroyd, and Jonathan D Cohen (2004). "ERP correlates of feedback and reward processing in the presence and absence of response choice". In: *Cerebral cortex* 15.5, pp. 535–544.

Yu, Rongjun and Xiaolin Zhou (2006). "Brain responses to outcomes of one's own and other's performance in a gambling task". In: *Neuroreport* 17.16, pp. 1747–1751.

6.3 A normative perspective on differential learning rates

Other studies with various protocols have also found differential learning associated with positive and negative outcomes (Frank, Seeberger, and O'reilly, 2004; Sharot, Korn, and Dolan, 2011; Niv et al., 2012; see Sharot and Garrett, 2016 for a review). This difference is often interpreted as a cognitive bias, or perhaps the result of limited cognitive resources, but this pervasive asymmetric updating actually raises normative questions. Cazé and van der Meer (2013) have recently tested the performance of agents able to differentially update positive and negative prediction errors, under action-outcome contingencies that are rarely tested in human participants.

6.3.1 Previous findings

Classical theories of reinforcement learning assume that action values are learnt via the calculation of a reward prediction error, i.e., the difference between the obtained and the expected outcome, independently of the valence of the prediction error. This way action values represent a weighted average of the past reward associated with each action, and of the initial estimate Q_0 (Sutton and Barto, 1998). Provided the learning rate is small enough and the environment is stationary, the action values will optimally converge to the average reward associated with that action. But using differential learning rates for positive and negative outcomes will cause the Q-value to be a biased estimate of the underlying average reward.

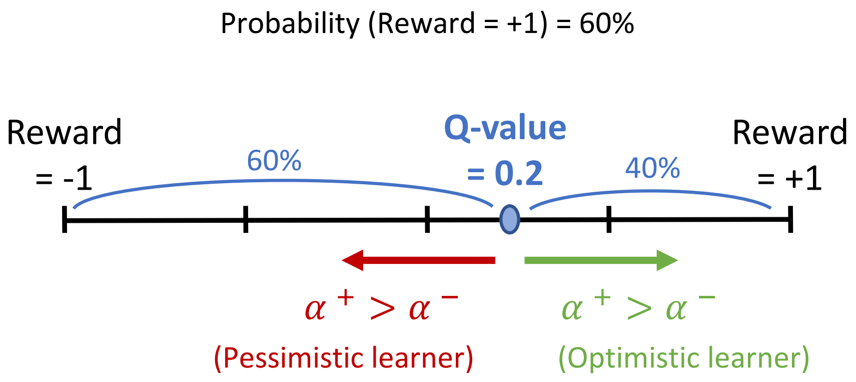


Figure 6.2: Q-values estimate the action average reward. Here we take the example of an action yielding positive outcomes (r = +1) in 60 per cent of cases, and thus negative outcomes (r = -1) in 40 per cent of cases. Asymmetric learning rates will cause the Q-value to be a biased estimate of average reward. For optimistic learners ($\alpha^+ > \alpha^-$), the Q-value will overestimate the true average reward, as positive outcomes will be preferentially updated. In contrast for pessimistic agents, the Q-value will underestimate the true average reward.

To obtain the Q-value distortion exerted by different α^+ and α^-, Cazé and van der Meer (2013) derived the equation for Q-value differential update for one action and one state. At a steady state, they found:

$$Q_\infty = \frac{px - (1-p)}{px + (1-p)} \tag{6.1}$$

where p is the probability of positive outcomes, and x is the ratio between the learning rate for positive over negative prediction error $x = \alpha^+/\alpha^-$.

Their simulations of Q-values after 800 trials for different values of p were consistent with the above equation. They found the Q-values of a rational learner ($\alpha^+ = \alpha^-$) to be good estimates of the reward average (here $2p - 1$). But for an optimistic agent ($\alpha^+ > \alpha^-$), the Q-values were an overestimation of the true average reward. In contrast a pessimistic agent ($\alpha^+ < \alpha^-$) underestimated the true average reward.

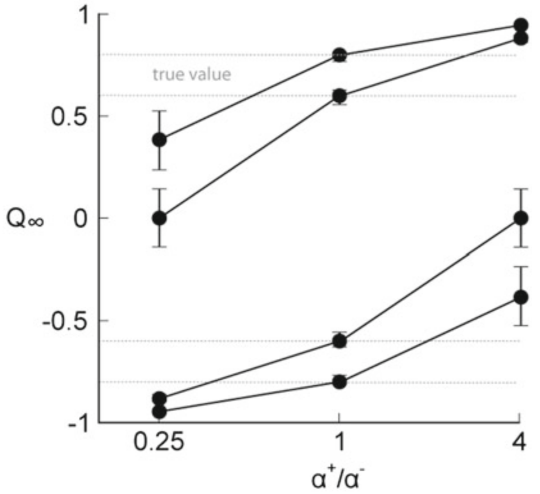


Figure 6.3: Differential learning rates result in biased estimates of the true expected values. Estimated Q-values after 800 trials averaged over 5,000 simulations for different ratios of α^+ and α^-. The dotted lines represent the true values of Q: 0.8, 0.6, -0.6, -0.8. The error bars represent the variance of the estimated Q-values. From the upper to the lower lines, the probability of positive outcome were respectively 0.9, 0.8, 0.2, 0.1. (Figure reproduced from Cazé and van der Meer, 2013).

To perform well, a model has to maximize the difference of Q-values between the high- and the low-rewarding actions. It is interesting to see that when the probabilities were 0.9 or 0.8, the model that maximized the difference between Q-values was the pessimistic learner, while the Q-values for $p = 0.2$ or $p = 0.1$ were the most divergent for an optimistic learner. From the previous equation, Cazé and van der Meer (2013) had computed the ratio for which ΔQ_∞ is maximal:

$$x^* = \frac{\sqrt{q_0 q_1}}{\sqrt{p_0 p_1}} \tag{6.2}$$

where p_1 is the probability of positive outcome for one action, p_0 is the probability of positive outcome for the other action, and q is the probability of negative outcome for each action: $q_i = 1 - p_i$.

We can see that, when both action yield close outcome probabilities ($p_1 \to p_0$), the optimal ratio between the positive over negative learning rate tends to be the ratio between p(negative reward) and p(positive reward). Therefore behavior is optimal when the positive (resp. negative) learning rate corresponds to the probability of negative (resp. positive) outcome. Among other models, Cazé and van der Meer (2013) have simulated a meta-learner, which adapts its learning rates accordingly to the tasks' underlying reward probabilities.

The authors first compared the rational, pessimistic and optimistic learners on two "two-armed bandit" tasks:

- a "low-reward" task, in which the probabilities of positive outcome were 0.2 and 0.1 for the two possible actions, therefore outcomes were mostly negative.
- a "high-reward" task, with probabilities 0.9 and 0.8, thus yielding mainly positive outcomes.

The model with the greatest difference in Q-values in the previous figure outperformed the other models. Indeed, in the low-reward task, the optimistic agent learnt to take the best action significantly more often than the rational agent, which in turn performs better than the pessimistic agent. In contrast, for the high-reward task, the pessimistic model outperformed the optimistic one. This decrease in performance can be explained by excessive exploration, as the models with lower performance also had high probabilities of switching actions.

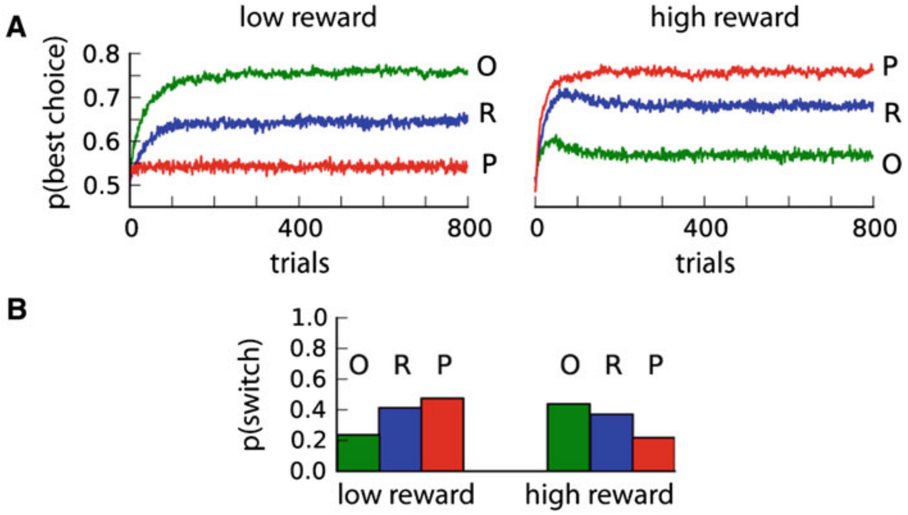


Figure 6.4: A. Performance for the three agents: "rational" (R, $\alpha^+ = \alpha^-$, blue line), "optimistic" (O, $\alpha^+ > \alpha^-$, green line) and "pessimistic" (P, $\alpha^+ < \alpha^-$, red line) in the low-reward (left panel) and high-reward (right panel) tasks. **B.** Proportion of action switch after 800 trials for each agent, in the two different tasks. (Figure reproduced from Cazé and van der Meer, 2013).

As different patterns of α^+ and α^- can only be advantageous in one of the two tasks, a meta-learner who could outperform a rational learner on both tasks was created. The meta-learner model had a ratio between the positive and negative learning rate that would optimally tend to the ratio between p(negative reward) and p(positive reward). It outperformed a rational learner in both low- and high-reward tasks, for a wide range of rational learning rate ($\alpha = 0.01$, 0.1 or 0.4).

Finally the different models were simulated on a task whose probabilities of positive outcome were 0.25 and 0.75, which is closer to the experiments in human reinforcement-learning. This time, different learning rates for positive and negative outcomes were not advantageous, as a rational learner ($\alpha^+ = \alpha^-$) outperformed all other models. But the authors' previous findings did generalize well to the case of a "three-armed bandit" task.

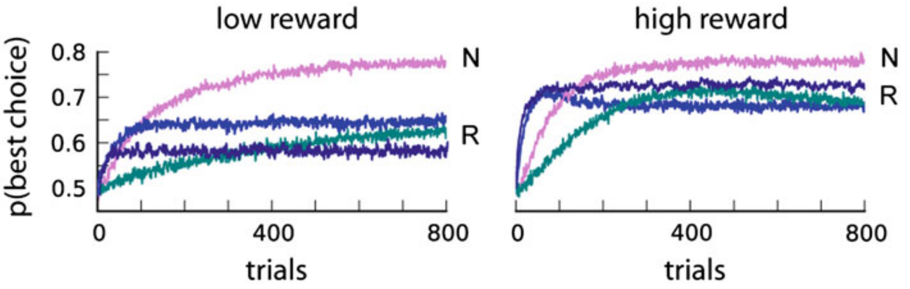


Figure 6.5: Performance of a meta-learner (N, in purple) and rational agents (in teal for $\alpha = 0.01$, in royal blue for $\alpha = 0.1$ and in navy blue for $\alpha = 0.4$) in the low-reward (left panel) and high-reward (right panel) tasks. (Figure reproduced from Cazé and van der Meer, 2013).

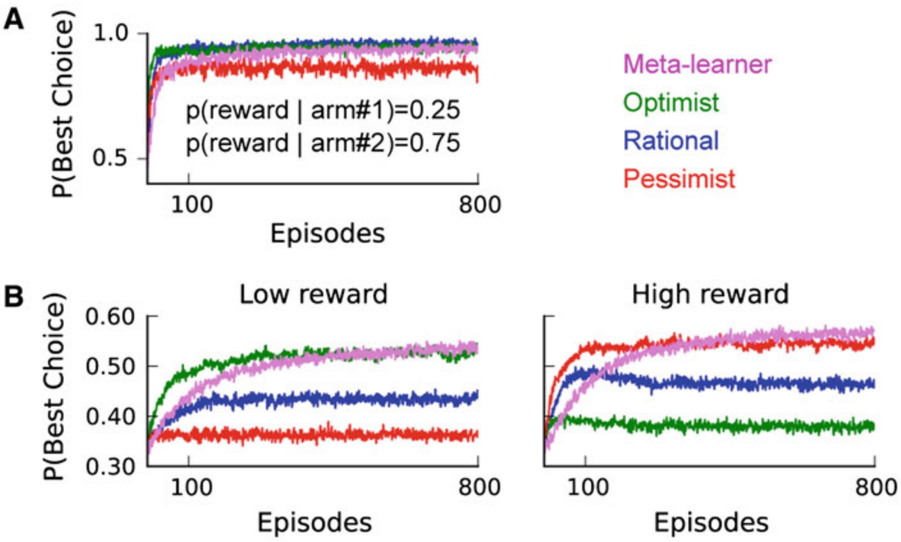


Figure 6.6: **A.** The performance of the different agents (Meta-learner, Optimistic, Rational and Pessimistic) in a task where the probabilities of reward are 0.75 and 0.25 for the two choices. **B.** The performance of agents in a "three-armed bandit" task, with reward probabilities 0.2, 0.15 and 0.1 in the low-reward task, and 0.9, 0.85 and 0.8 in the high-reward task. (Figure reproduced from Cazé and van der Meer, 2013).

6.3.2 Our published replication article

ReScience

[Re] Adaptive properties of differential learning rates for positive and negative outcomes

Sophie Bavard[1] **and Héloïse Théro**[1]

1 Laboratoire de Neurosciences Cognitives Computationnelles (ENS - INSERM), Département d'Études Cognitives, École Normale Supérieure, PSL Research University, 29 rue d'Ulm, 75005 Paris, France

sophie.bavard@gmail.com, thero.heloise@gmail.com

Editor
Olivia Guest

Reviewers
Xavier Hinaut
Benoît Girard

Received Feb, 20, 2018
Accepted Jun, 14, 2018
Published Jun, 14, 2018

Licence CC-BY

Competing Interests:
The authors have declared that no competing interests exist.

Article repository

Code repository

A reference implementation of

→ Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. Biological cybernetics, 107(6), 711-719. https://doi.org/10.1007/s00422-013-0571-5

Introduction

Reinforcement learning represents a fundamental cognitive process: learning by trial and error to maximize rewards and minimize punishments. Current and most influential theoretical models of reinforcement learning assume a unique learning rate parameter, independently of the outcome valence (Sutton and Barto [14], O'Doherty et al. [10], Behrens et al. [1]). However human participants were shown to integrate differently positive and negative outcomes (Frank, Seeberger, and O'Reilly [3], Frank et al. [4], Sharot, Korn, and Dolan [13]). This motivated the reference article to implement a modified version of the reinforcement learning model, with two distinct learning rates for positive and negative outcomes (Cazé and Meer [2]).

They have shown that although differential learning rates shifted reward predictions and could thus be seen as a maladaptive bias, this model can outperform the classical reinforcement learning model on tasks with specific outcome probabilities. Following Cazé and Meer [2]'s predictions, a subsequent empirical article have modeled human behavior on these specific tasks (Gershman [7]). The question is still an active research area, as various articles have further investigated the difference learning rates bias (Garrett and Sharot [5], Moutsiana et al. [9], Shah et al. [12], Garrett and Sharot [6], Lefebvre et al. [8], Palminteri et al. [11]).

A link to the pdf version of the reference article was posted on the last author's laboratory website (http://www.vandermeerlab.org/publications.html), but the corresponding code was not available (https://github.com/vandermeerlab/papers/tree/master/Caze_vanderMeer_2013). We believe that an openly available code repository replicating the results of Cazé and Meer [2]'s paper can be helpful to the scientific community. We therefore implemented the model and analysis scripts using Python, with numpy, random and matplotlib libraries.

Methods

We first implemented our scripts on Matlab, as we were more familiar with this language, and then adapted them on Python.

We used the modeling description of the reference article to implement our replication. They used standard Q-learners with a softmax action selection rule (Sutton and Barto [14]), and their precise description enabled us to implement them with low difficulty. But we found four ambiguities in the simulation procedure.

First, the authors described their analytical results to be valid for "$Q_0 \neq \{-1, 1\}$" in section 2, but did not specify what value of Q_0 they used in all the following simulations. We chose to use $Q_0 = 0$, as this initial value is the middle point between the two possible outcomes (i.e., -1 and 1). As we replicated all the original figures, even the dynamics in the beginning of the learning curves (see Figures 2 A, 3 and 4 B), we believe the reference article must have used similar initial Q-values.

Second, regarding the parameter setting for Figure 1's simulations, the ratio of α^+ over α^- was said to be either 0.25, 1 or 4, but they did not specify what were the exact values of α^+ and α^- used. We thus set them according to the following description of the pessimistic, rational and optimistic agents in section 3, i.e.,:

- $\alpha^+ = 0.1$ and $\alpha^- = 0.4$ for the ratio of 0.25
- $\alpha^+ = 0.1$ and $\alpha^- = 0.1$ for the ratio of 1
- $\alpha^+ = 0.4$ and $\alpha^- = 0.1$ for the ratio of 4

Third, the number of iterations made to generate Figures 3 and 4 were not indicated, and we assumed the authors used the same number as in Figures 1 and 2 (i.e., 5,000 runs).

Finally, in the reinforcement learning framework, the probabilities to choose each action are computed, then used to select an action through a pseudo-random generator. In the reference article, it was sometimes unclear whether the analyses were performed on the probabilities of choice, or rather the proportions of implemented choices. For example Figure 2's legend indicated: "Mean probability of choosing the best arm", suggesting that the probabilities themselves were used. However, when commenting the figure in section 3, the authors appeared to say that the actual choices were rather used: "the optimistic agent learns to take the best action significantly more than the rational agent". For our analyses, we started by using the probabilities of choice, as this would lead to more clear, less noise-corrupted results. However we then obtained very smooth learning curves, and were unable to reproduce the spikiness of the original Figures 2, 3 and 4. We thus computed the proportions of implemented choices for all our figures.

Results

We numbered our figures in the same way as the reference article.

All our figures reproduced the patterns of the original results. We were even able to replicate the fine-grained details of the learning curves, like the early bumps in performance in the high-reward task (Figures 2 A, 3 and 4 B, right panels, around 50-100 trials). In Figure 1, the mean and the variance of the Q-values were also very similar as the ones in the original figure.

The only discrepancy we found was in Figure 4 A. Although the general pattern was replicated, our learning curves appeared smoother than in the reference article. As the number of simulations were not explicitly specified for this figure, we cannot know if this is due to us running a higher number of simulations than the reference article, or from another difference in model implementation.

Conclusion

All the figures in Cazé and Meer [2] have been successfully reproduced with high fidelity, and we confirm the validity of their simulations. Overall the whole replication

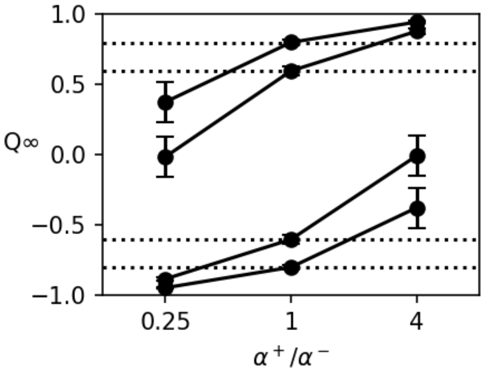


Figure 1: Average estimated Q-values after 800 trials averaged for different ratios of α^+ and α^-. The dotted lines represent the underlying average reward: 0.8, 0.6, -0.6, -0.8. The error bars represent the variance of the estimated Q-values.

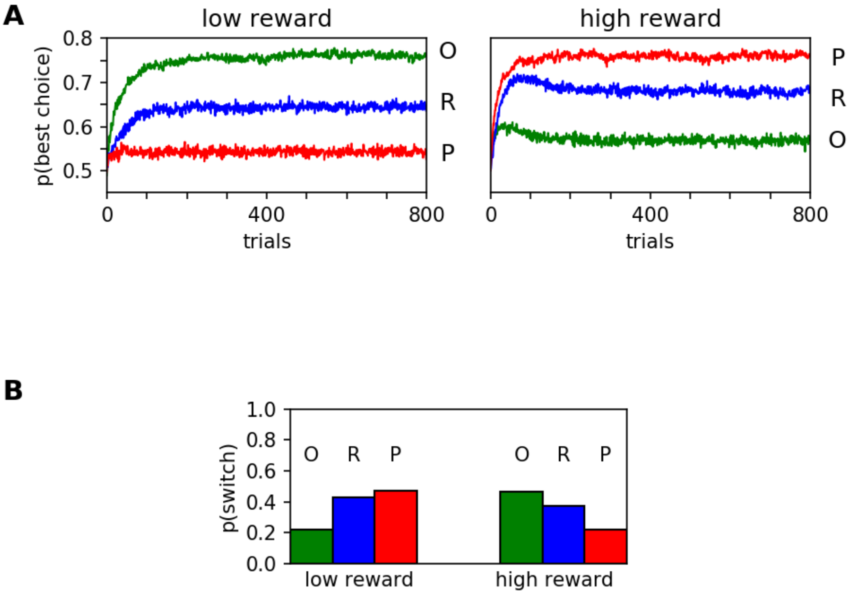


Figure 2: **A.** Performance, i.e. proportion of choices for the best action, for the three agents: Rational (R, $\alpha^+ = \alpha^-$, blue line), Optimistic (O, $\alpha^+ > \alpha^-$, green line) and Pessimistic (P, $\alpha^+ < \alpha^-$, red line). In this figure and the following ones, the left (resp. right) panel corresponds to the low-reward (resp. high-reward) task. **B.** Proportion of action switch after 800 trials for each agent, in the two different tasks.

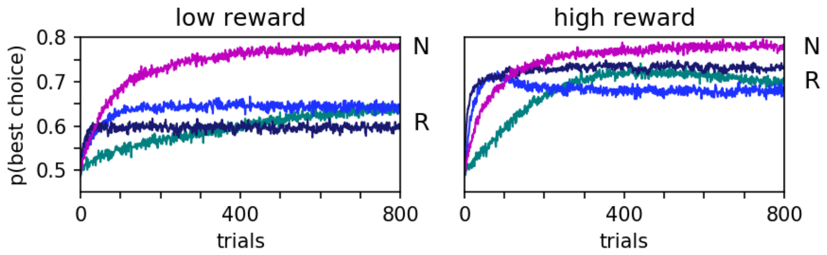


Figure 3: The performances of the Meta-learner (N) are shown in *purple* and those of the Rational agents (R) in different colors of blue (in *teal* for $\alpha = 0.01$, in *royal blue* for $\alpha = 0.1$ and in *navy blue* for $\alpha = 0.4$).

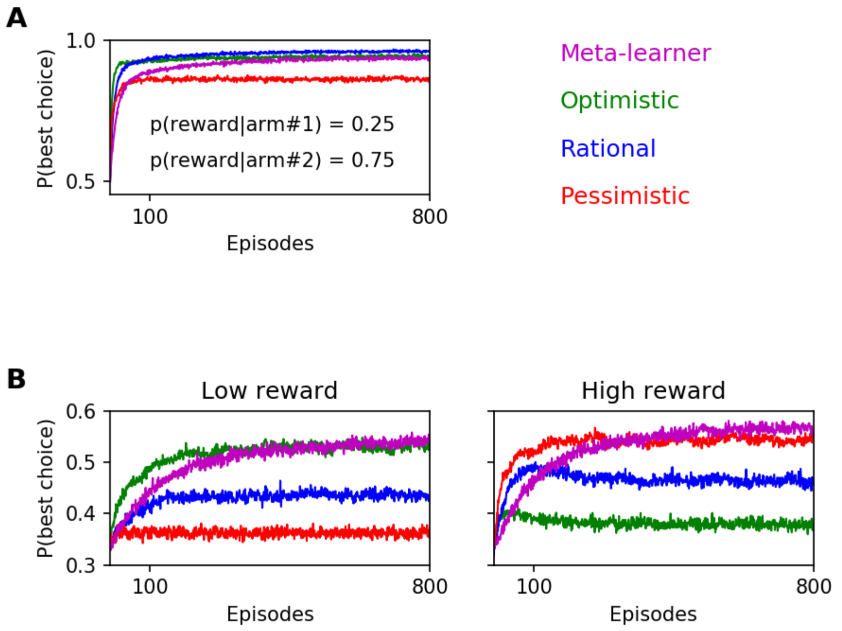


Figure 4: The performances of the Meta-learner, Optimistic, Rational and Pessimistic agents **A.** in a task where the probabilities of reward are 0.75 and 0.25 for the two choices. **B.** in a "three-armed bandit" task.

procedure was smooth: the models were implemented with low difficulty, and the simulations were quite straightforward apart from a few obscure details. We hope this replication can foster future research in the domain.

References

[1] Timothy EJ Behrens et al. "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9 (2007), p. 1214.

[2] Romain D Cazé and Matthijs AA van der Meer. "Adaptive properties of differential learning rates for positive and negative outcomes". In: *Biological cybernetics* 107.6 (2013), pp. 711–719.

[3] Michael J Frank, Lauren C Seeberger, and Randall C O'Reilly. "By carrot or by stick: cognitive reinforcement learning in parkinsonism". In: *Science* 306.5703 (2004), pp. 1940–1943.

[4] Michael J Frank et al. "Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning". In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16311–16316.

[5] Neil Garrett and Tali Sharot. "How robust is the optimistic update bias for estimating self-risk and population base rates?" In: *PLoS One* 9.6 (2014), e98848.

[6] Neil Garrett and Tali Sharot. "Optimistic update bias holds firm: Three tests of robustness following Shah et al." In: *Consciousness and cognition* 50 (2017), pp. 12–22.

[7] Samuel J Gershman. "Do learning rates adapt to the distribution of rewards?" In: *Psychonomic bulletin & review* 22.5 (2015), pp. 1320–1327.

[8] Germain Lefebvre et al. "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4 (2017), p. 0067.

[9] Christina Moutsiana et al. "Human frontal–subcortical circuit and asymmetric belief updating". In: *Journal of Neuroscience* 35.42 (2015), pp. 14077–14085.

[10] John O'Doherty et al. "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". In: *science* 304.5669 (2004), pp. 452–454.

[11] Stefano Palminteri et al. "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8 (2017), e1005684.

[12] Punit Shah et al. "A pessimistic view of optimistic belief updating". In: *Cognitive Psychology* 90 (2016), pp. 71–127.

[13] Tali Sharot, Christoph W Korn, and Raymond J Dolan. "How unrealistic optimism is maintained in the face of reality". In: *Nature neuroscience* 14.11 (2011), p. 1475.

[14] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge, 1998.

6.3.3 Further analyses of the Study II's data

We will show here that this normative interpretation for positive and negative learning rates cannot explain our results in Study II. In this study, we used symmetrical action-outcome contingencies (0.8 and 0.2), and we found the participants' positive learning rates to be higher than the negative ones.

We have simulated the optimistic, rational and pessimistic models described in Cazé and van der Meer (2013) to see what performances these models would have had in the two experiments published in the Quartely Journal of Experimental Psychology. We only adapted these models by normalizing their Q-values. We ran 100 simulations on the design matrices used for each participant. We can see that the model with the highest performance was the pessimistic learner, although our participants' best-fitting parameters showed they were optimistic.

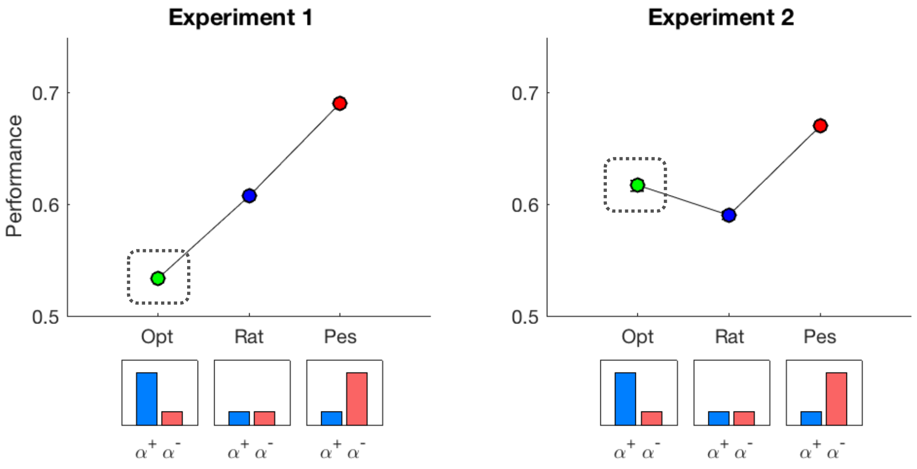


Figure 6.7: The performance of the different agents (Optimistic, Rational and Pessimistic) in our two experiments published in the Quartely Journal of Experimental Psychology (Study II). We have circled the performance corresponding to the participants' pattern of learning rates.

Cazé and van der Meer (2013) also tested their models on contingencies close to the ones we used (.75 and 0.25, see Figure 6.6A), but they found that : "In this scenario the advantage of differential learning rates is negligible". An important difference is that Cazé and van der Meer (2013) tested their models in stationary settings, in which the action-outcome contingencies were stable during the whole experiment, while we used a reversal-learning procedure in Study II.

Palminteri et al. (2017) did investigate the behavior of optimistic and rational models when action-outcome contingencies reversed. They found that the optimistic model was slower to inverse its values after a reversal, therefore displaying worse performance. They also found the participants with a higher optimistic bias to perform less well after a reversal than the participants having a lower or nonexistent bias. This shows that being optimistic is not optimal in a reversal-learning setting such as the one we used in Study II. A normative perspective thus cannot explain why we found our participants to have a higher positive, than negative, learning rate ($\alpha^+ > \alpha^-$).

6.3.4 Further analyses of the Study III's data

Cazé and van der Meer (2013)'s theory is that different learning rate patterns are adaptive in low- and high-reward environment. As we have shown in the draft, we could see in both experiments that optimistic models had indeed better performances in the low-reward condition (0.4 and 0.1 contingencies), and that pessimistic models were optimal in the high-reward condition (0.6 and 0.9 contingencies, see Figure 5 of the draft).

We wanted to test if our participants were able to adapt their learning rates according to which condition they were in, to increase their performance. When we fitted the high- and low-reward conditions separately, we found no clear and replicable differences in learning rates, although it seemed that participants were less optimistic in the low- than high-reward condition. We have thus displayed in Study III only pooled results (Figure 3 of the draft).

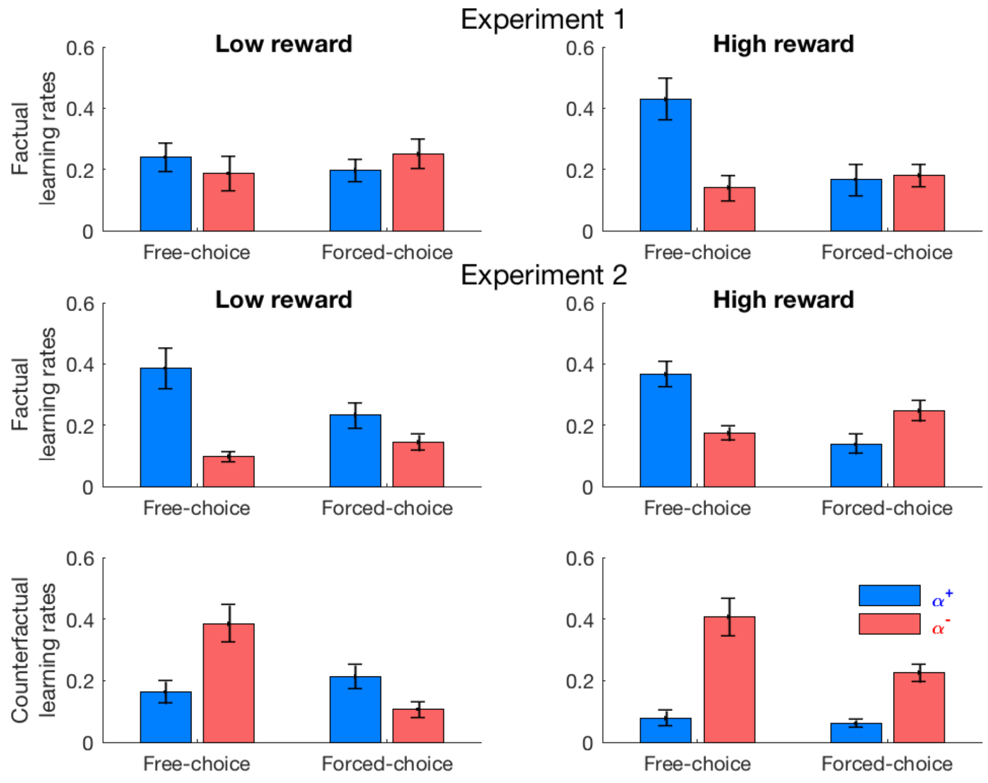


Figure 6.8: The differential learning rate pattern when we fitted separately the low- and high-reward conditions in Study III.

6.3.5 Reanalyses of Gershman (2015)'s data

Gershman (2015) has also tested low- and high-reward contingencies on human participants, and found no difference in learning rates between the conditions. We contacted him, and he kindly sent us his data. We reanalyzed them, and ensured to center the initial Q-values with respect to the outcome distribution (i.e., the Q-values were initialized at 0.5 as the outcomes were either +1 or 0). One should note that here only factual learning rates are fitted, as the participants only saw factual outcomes.

Except in Experiment 1's results, we found participants' best-fitting learning rates to be slightly different in low- and high reward contingencies: participants appeared to be more pessimistic in low-reward conditions, and more optimistic in high-reward conditions. This effect was small and non-existent in the first experiment, and it was also not very clear in Study III. We thus think that further research is needed before we can conclude whether participants adapt their learning rates in low- and high-reward conditions.

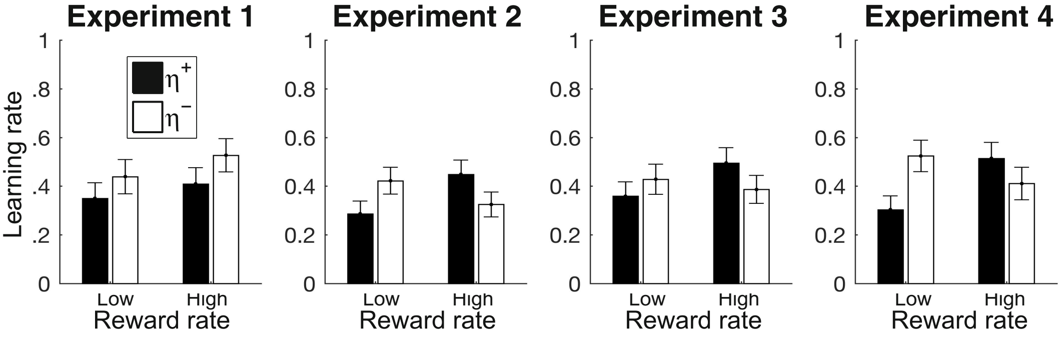


Figure 6.9: Our reanalysis of Gershman (2015)'s data.

It should be noted that this effect, if it is confirmed, goes in the opposite direction than what a normative approach would recommend. Indeed the optimal meta-learner developed by Cazé and van der Meer (2013) was actually more optimistic in low-reward conditions, and more pessimistic in high-reward conditions. We can thus interpret Gershman (2015)'s participants as displaying a frequency effect: people appeared to integrate more outcomes that are frequently seen. By contrast the optimal learning process is to learn more from rare outcomes, as they are the most informative.

Differential learning rates for positive and negative outcomes can be advantageous in some experimental settings. Indeed, Cazé and van der Meer (2013) found that a higher positive than negative learning rate allowed for better performances in a low-reward environment (i.e., when both actions had low probabilities of positive outcomes) while the inverse pattern can be found in a high-reward environment. Our replication article confirmed the validity of these results.

It should be noted that this normative perspective was in contrast with most of our results. Indeed the optimistic learning rate pattern we found in Study II was actually sub-optimal in a reversal-learning environment. By reanalyzing Study III's and Gershman (2015)'s data, we also found that people, if anything, seem to adapt their learning rates in the opposite direction than what optimality would recommend.

Chapter 7

General Discussion

In this PhD thesis, we have used cognitive modeling to investigate the relationship between control, agency and reinforcement learning in human decision-making.

In Study I, a series of 3 experiments were built on a modified reversal-learning procedure, in which there was some uncertainty about the identity of the causal agent. There were different conditions in which the participant's actual control over the outcomes could be positive or null. Through model comparison, we found that the model best able to fit and simulate participants' behavior was not a model explicitly looking for control, but rather a model based on counterfactual emulation, i.e., the model's choice is assumed to always control the action outcome. Moreover, this counterfactual emulation was hierarchically implemented at the different action level, suggesting a hierarchical representation of the possible actions in the participants' mind. This hierarchical counterfactual emulation was found in all conditions, regardless of the actual instrumental control implemented.

In Study II, we also used a reversal-learning paradigm while measuring intentional binding, a proxy to the implicit feeling of agency. We were interested in the fluctuation of sense of agency that accompanies adaptive behavior. We observed in 3 experiments a post-error boost of action binding: action binding on the trial following a non-rewarded outcome was stronger than following a rewarded outcome. Interestingly, we found participants' best-fitting learning rates to be higher for positive than negative outcomes ($\alpha^{+} > \alpha^{-}$), and the post-error boost was inter-individually correlated with the asymmetry in learning rates. Besides our classical 'learning' condition, we also implemented a 'random' condition, in which participants were explicitly instructed that action-outcome mappings were entirely unpredictable. We found the post-error boost of action binding to be specific to a learning context. It should be noted that our best-fitting model was a normalized reinforcement learning model, equivalent to the counterfactual emulation model described in Study I, with no hierarchy between the chosen and unchosen actions ("flat" counterfactual emulation: $\alpha_{CF} = \alpha$).

Finally, in Study III, we conducted two stationary instrumental conditioning tasks to investigate reinforcement learning processes occurring when the participant' choice was either free or forced. Previous experiments have shown that people usually display a choice-confirmation bias, i.e., they preferentially take into account information that confirms their current decision ($\alpha_{F}^{+} > \alpha_{F}^{-}$ and $\alpha_{CF}^{+} < \alpha_{CF}^{-}$). We replicated this result in free-choice trials, and found that, when participants were forced to match a preselected option, they integrated outcomes independently from their valence ($\alpha_F^+ = \alpha_F^-$ and $\alpha_{CF}^+ = \alpha_{CF}^-$). Interestingly, Cazé and van der Meer (2013) have shown in silico that different learning rates can be advantageous in certain experimental contingencies. We replicated this article, and used similar simulations to address the optimality of our different findings.

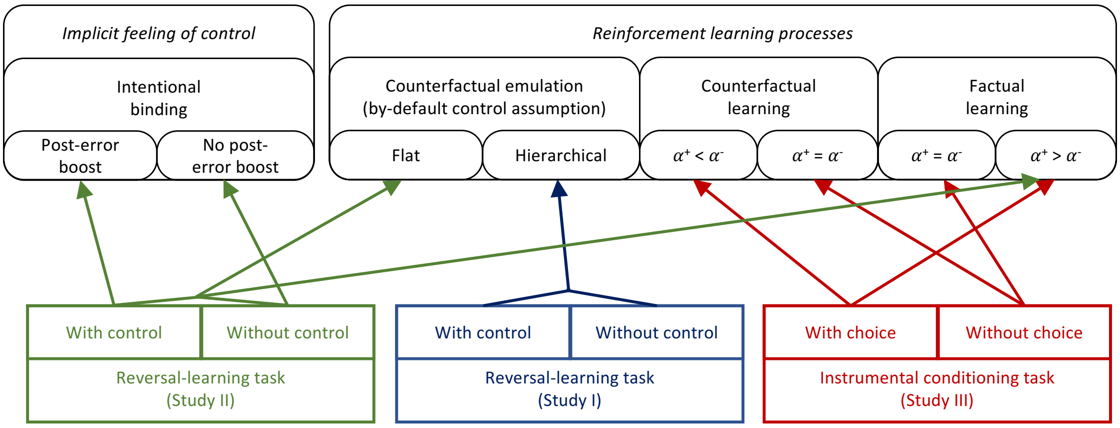


Figure 7.1: A summary of our findings.

Our general conclusion is that control perception and reinforcement learning, two fundamental fields of human psychology, are deeply intertwined. Indeed, action binding, an implicit proxy for the feeling of agency is influenced by post-error adaptive mechanisms. Furthermore, contrary to impartial machines, humans care about being in control or about making the right choice, and this results in integrating information in a one-sided way.

In this discussion, we will now try to parallel these results with certain influential cognitive biases and psychological traits. Then we will see how our results can help formalizing the notion of control, to characterize what is grounding people's perception of control.

7.1 Cognitive biases and psychological traits

First, we will discuss our findings with respect to some cognitive biases and psychological traits that have been observed in behavior or explicit reports. We have seen in the introduction that they was a gap between behavior and explicit reports so it should be noted that generalizing our findings to phenomenological experiences would be an extrapolation.

7.1.1 Valence-induced biases

One of the most salient attributes of information is valence: whether a piece of news is good or bad. Most of the classic theories assume that agents gather and integrate information in a manner that will result in a relatively accurate representation of reality. But examining people's beliefs about themselves and their future reveals systematic biases. In approximately 80% of the population, desirable information is integrated into prior beliefs more readily than undesirable information, resulting in an optimism bias (Sharot and Garrett, 2016).

The optimism bias may be counterintuitive, as most people would say they remember more vividly negative than positive events. Indeed, a general negativity bias was also found in different experiments, based on both innate predispositions and experience to give greater weight to negative entities (e.g., events, objects, personal traits). In an influential review, Rozin and Royzman (2001) concluded that there was a pervasive negativity bias, that could actually be meaningful and adaptive, in much of human and animal cognition and behavior.

In this PhD thesis, we found differences in learning rates for positive and negative outcomes. However we cannot conclude in favor of a general positivity or negativity bias, as this difference in learning rates took various forms. Indeed we found higher positive than negative learning rates for factual outcomes (Studies II and III), but the reversed pattern for counterfactual outcomes (Study III), and no valence-induced difference in forced-choice trials (Study III). Although the negativity bias can be useful to understand how humans process external information, we would argue that human behavior in a reinforcement learning context is better explained by self-related biases, rather than a general valence-induced bias.

7.1.2 The cognitive dissonance theory

In Aesop's Fable "The Fox and the Grapes", a fox tries to get some grapes that are hanging on a high, unreachable vine. After failing to reach them, the fox decides that the grapes were probably sour anyway. An interesting aspect of this story is the idea that making a choice (e.g., giving up on the grapes) can change one's preferences.

In a seminal study, 225 female students rated a series of domestic appliances and then were asked to choose among two equally preferred appliances as a gift. The results of a second round of ratings indicated that the students increased their ratings of the domestic appliance they had selected as a gift and decreased their ratings of the appliances they had rejected (Brehm, 1956). Young children and non-human primates were also shown to exhibit choice-induced preferences (Egan, Bloom, and Santos, 2010).

This paradigm was originally developed in order to study the phenomenon of cognitive dissonance reduction. According to this theory, the action of deciding provokes a psychological dissonance consequent to choosing X instead of Y, despite little difference between X and Y. Thus, the decision "I chose X" is dissonant with the cognition that "There are some aspects of Y that I like.". People would then artificially inflate their preference to X and decrease their preference for Y to reduce the cognitive dissonance.

The dissonance theory has been generalized to also include inconsistency between two cognitions, and not only between cognition and action (Festinger, 1957). Dissonance theory is more than simply a theory about consistency. It is essentially a theory about sense-making: how people try to make sense out of their beliefs, their environment, and their behavior – and thus try to lead lives that are (in their own minds) reasonable, sensible, and meaningful (Aronson, 1997).

We found in Studies II and III that participants preferentially took into account information that confirms their decision, except when their choice was not intentional, but imposed by the "computer", i.e., an external source. These results are consistent with the cognitive dissonance theory, and more specifically with choice-induced preference. Indeed if people integrated more information consistent with their choice, this biased learning process would lead to a choice-induced preference behavior. This was shown in silico by Lefebvre et al. (2017): after a series of choices, a model with a higher positive than negative learning rate displayed a pronounced preference for one option, although both options actually were equally rewarding.

By defining dissonance as a negative drive state, Leon Festinger combined motivation with cognition and formulated new predictions that could not be easily explained by other theories. For example, reinforcement theory would suggest that, if you reward individuals for making a particular statement, they might come to like and believe in the truth and beauty of that statement (through the mechanism of secondary reinforcement). But Festinger and Carlsmith (1959) actually showed the opposite result. Participants were subjected to a boring experience and then paid either $1 or $20 to tell someone that the experience had been interesting and enjoyable. The participants who said that they found the task enjoyable in order to earn $1 came to actually believe it was enjoyable to a far greater extent than those who were paid $20 to lie.

Cognitive dissonance theory was often used to explain illogical, or even disadvantageous, behavior. Interestingly, in Study III, we found choice-confirmatory models to outperform valence-neutral models. It was non trivial to see that a unique learning model can both maximize rewards and minimize cognitive dissonance under some experimental conditions.

It should be noted that another recent cognitive model is also compatible with the cognitive dissonance theory: the self-consistent Bayesian observer model. This model made for perceptual decision-making assumes that a subject will integrate sensory evidence in a manner that is consistent with the subject's preceding choice (Luu and Stocker, 2018). We hope further cognitive modeling approaches will soon be able to explain the fine-grained details of cognitive dissonance mechanisms.

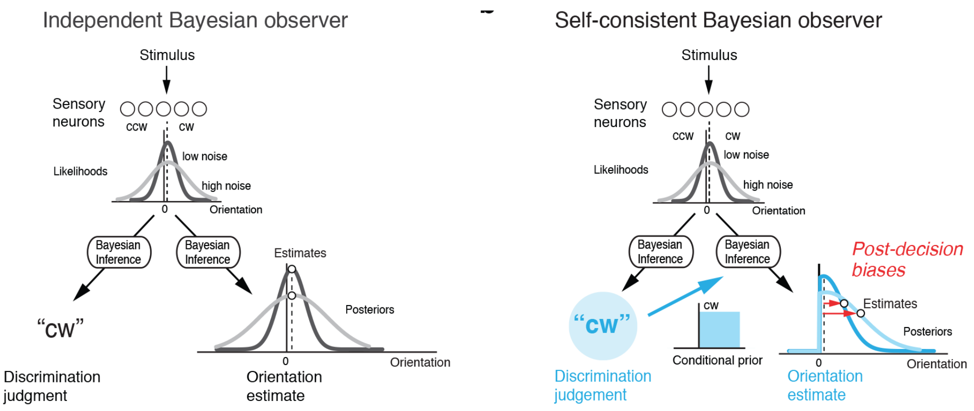


Figure 7.2: The self-consistent Bayesian model (shown on the right panel) is used to explain why two successive choices made by participants were more consistent than what is predicted by a normative model making two independent decisions (shown on the left panel). (Figure reproduced from Luu and Stocker, 2018).

7.1.3 The self-serving bias

Most people rate their abilities as better than 'average' even though it is statistically impossible for most people to have better-than-median abilities. In a survey of faculty at the University of Nebraska-Lincoln, 68% of professors rated themselves in the top 25% for teaching ability, and more than 90% rated themselves as above average (Cross, 1977). High school students ascribed higher levels of honesty, persistence and originality to themselves than to the average student, and also described themselves as less hostile, less vain and less unreasonable than average. The relative over-evaluation of one's own attributes has been shown in such diverse domains as personality traits, abilities and satisfaction with relationships (Hoorens, 1993).

This 'above-average' bias, conjugated with cognitive dissonance, may explain the self-serving bias, occurring when people make internal attributions for desired outcomes and external attributions for undesired outcomes. This bias is evident in workers who attribute receiving promotions to hard work and exceptional skill, yet attribute denial of promotions to unfair bosses, and in drivers who attribute accidents to the weather or other drivers, yet attribute the narrow avoidance of an accident to their alertness and finely honed driving skills (Shepperd, Malone, and Sweeny, 2008).

Interestingly, the self-serving bias is in contradiction with the post-error boost of action binding we found in Study II. Because people tend to attribute more negative outcomes to external factors, they should feel *less* agent after a negative outcome. Yet we found higher action binding following a non-rewarding tone, than following a rewarding tone. As action binding is supposed to reflect an implicit, maybe pre- or non-conscious, form of agency, it is possible that action binding is not subjected to the self-serving bias. However, Takahata et al. (2012) found intentional binding to be attenuated by negative monetary outcome, consistently with the self-serving bias.

A crucial detail in Takahata et al. (2012) experiment is that only one key could be pressed by the participant, excluding any possibility for adaptive behavior to emerge. In Study II tasks, two possible actions could be chosen, and negative out-

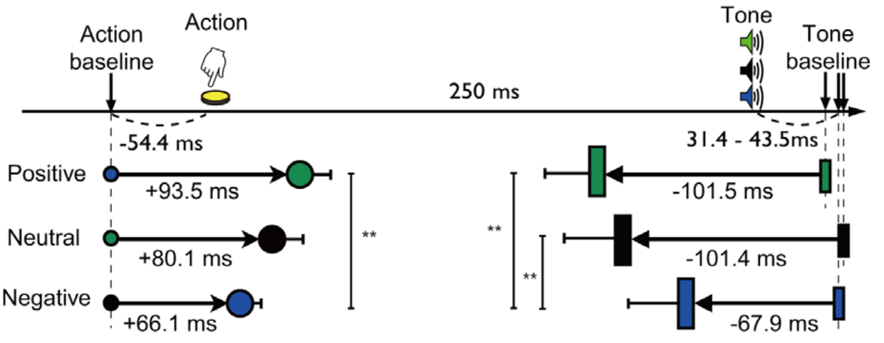


Figure 7.3: Action and outcome binding for auditory stimuli paired with positive, neutral or negative monetary outcomes and following a key press. (Figure reproduced from Takahata et al., 2012).

comes are known to be motivationally salient events triggering the necessity of adaptation (Wessel et al., 2014). We would thus postulate that the self-serving bias does not arise in adaptive behavior tasks, in which error processing is crucial to increase performance.

People were found to show post-error adaptations, potentially to improve their performance in the near future. At least three types of behavioral post-error adjustments have been observed: post-error slowing, post-error reduction of interference, and post-error improvement in accuracy, as well as neuronal activity increase in task-relevant brain areas, and activity decrease in distracter-encoding brain areas (Danielmeier and Ullsperger, 2011).

The general increase in attention and vigilance following an error may be the cause for an increase in the feeling of agency. An interesting perspective to Study II would then be to study the relationship of the post-error boost of action binding, task performance, and error awareness, as it is still unclear which post-error adjustments actually depend on error awareness or even 'task difficulty awareness' (Ullsperger et al., 2010).

7.1.4 The need for control hypothesis

Superstitious and paranormal beliefs are widespread in the population and thus have attracted a great deal of attention from research. An acute state of anxiety correlates with paranormal beliefs (Keinan, 2002). Moreover, Dudley (1999) assessed the level of superstitious belief both before and after working on a solvable or unsolvable puzzle. Reported level of superstitious belief increased following exposure to unsolvable, but not solvable problems. It suggests that participants invoke superstitious beliefs during instances of uncontrollability. Paranormal believers also tend to be perceived by independent judges as trying to control others' actions.

Given that paranormal belief is related to fantasy proneness, its origins may be found in one of the antecedent factor of fantasy proneness, namely a history of abuse in childhood. Irwin (1992) has found a link between paranormal belief and childhood trauma, particularly physical abuse by family members. Traumatic events pose a potential threat to a state of assurance, in essence because they can be taken to imply that the world is uncertain and chaotic. By incorporating a system of personal beliefs, the individual has a cognitive framework for effectively structuring events and experiences in life, so that they can be mastered, at least intellectually (Irwin and Watt, 2007).

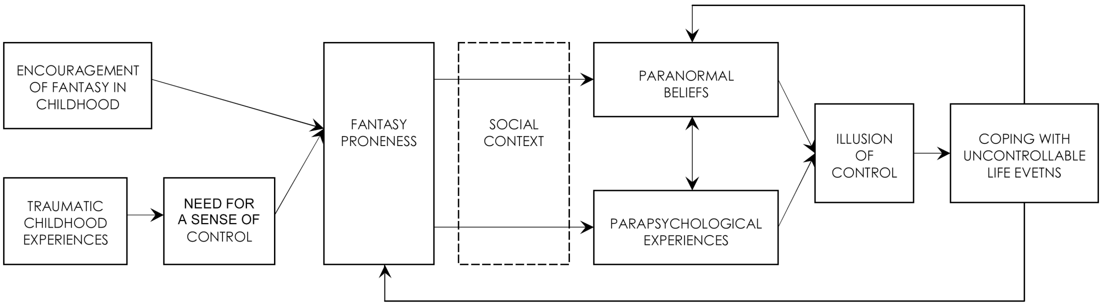

ENCOURAGEMENT OF FANTASY IN CHILDHOOD
TRAUMATIC CHILDHOOD EXPERIENCES
NEED FOR A SENSE OF CONTROL
FANTASY PRONENESS
SOCIAL CONTEXT
PARANORMAL BELIEFS
PARAPSYCHOLOGICAL EXPERIENCES
ILLUSION OF CONTROL
COPING WITH UNCONTROLLABLE LIFE EVETNS

Figure 7.4: A model of origins and functions of paranormal beliefs. (Figure reproduced from Irwin, 1993).

Empirical research offers support for this view. Blackmore and Trościanko (1985) have shown that a group of paranormal believers had a greater sense of control over a computer task that did a group of nonbelievers, yet the two groups did not differ in their achieved control of the task. Rudski (2004) found that questionnaire measures of illusion of control were associated with paranormal belief, particularly superstitious and precognition beliefs, again suggesting that such beliefs might give a sense of control over otherwise unpredictable events. It should be stressed that the particular form of paranormal belief endorsed by the individual will depend greatly on the cultural and social environment.

Whitson and Galinsky (2008) have shown that increased pattern perception had a motivational basis by measuring the need for structure directly. They found that people experiencing a loss of control were more likely to develop superstitions, but also to perceive conspiracies, to see images in noise and to form illusory correlations. Many of these distortions are typically discussed as separate phenomena, but they can actually be regarded as specific cases of a more general misperception of randomness.

Scheibehenne, Wilke, and Todd (2011) found that most of their participants preferred to predict purely random sequences over those with moderate negative autocorrelation and thus missed the opportunity for above-chance payoff. However, there exist important individual differences with regard to how strongly people are prone to that misperception, and with regard to how much they give into that misperception and bet on it (Scheibehenne and Studer, 2014). For example, gamblers appeared to be more impulsive than community members, and it could explain why they are more willing to bet impulsively on perceived illusory patterns (Gaissmaier et al., 2016).

In practice, field studies using control interventions have shown that new perceived control could be particularly beneficial for people who believed they had little control. For instance, elderly people often experience an overall loss of actual control, due to reduced mobility, retirement from work and increasing health problems. When they were given new control opportunities, even minor ones such as being asked to take care of themselves or water a plant, they show renewed resilience in psychological and physical well-being (Langer and Rodin, 1976), and these positive effects were shown to last as long as 18 months later (Rodin and Langer, 1977).

All in all, it might be better to err on the side of too much perceived control. Beliefs are thought to be held because they serve significant psychodynamical needs of the individual, and they can achieve this function whether they are grounded in objective reality or are intrinsically illusory (Taylor and Brown, 1988). Many instances have been found in which it seems to be better to think you have control than not, even in the case of dire circumstances (Taylor, Wayment, and Collins, 1993). Our results in Study I suggest that people do rely on a by-default control hypothesis, although it can make them unable to determine clearly which of their actions were actually instrumental.

7.1.5 Free Will and cognitive traits

In a series of experiments, Alquist et al. (2015) have manipulated and measured belief in free will. They also measured participants' counterfactual thinking by asking them to reflect on a time they had hurt someone and counting the number of thoughts in which they imagined what could have gone differently. Belief in free will was associated with more counterfactual thinking, and particularly with an an increase in the generation of self and upward counterfactuals, which have been shown to be particularly useful for learning.

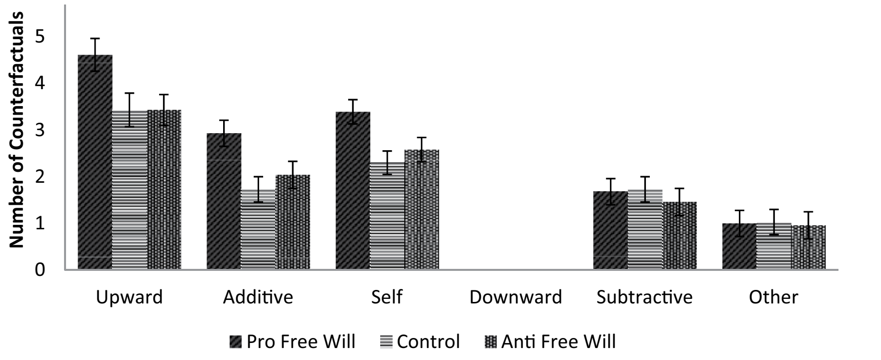


Figure 7.5: Average number of counterfactual thoughts by structure and free will condition. (Figure reproduced from Alquist et al., 2015).

These results parallel our findings in Study I, in which instrumental control is implemented as counterfactual power by participants. As Alquist et al. (2015), we also hypothesize that a belief in control can be beneficial for action-outcome learning, by triggering a counterfactual emulation mechanism.

In the introduction, we have seen how artificially decreasing belief in free will led to cheating and agressive behavior. However level of belief is stable in many cases, and personality psychologists had developed tools to quantify individual differences. To measure free will belief, researchers tend to rely on either the Free Will and Determinism Plus Scale (FAD+; Paulhus and Carey, 2011) or the Free Will and Determinism Scale (FWDS; Rakos et al., 2008). Other measures, including one-item or two-item questions about whether one believes in free will, are also sometimes used. One advantage of the FAD+ is that it measures free will belief and deterministic beliefs with separate scales. In contrast, the FWDS treats determinism as the polar opposite of free will, such that increases in one belief necessarily reflect decreases in the other (Baumeister and Brewer, 2012).

Questionnaire results were consistent with free will manipulation experiments. For example, students with higher dispositional belief in free will reported greater expectations of future professional success. This significant prediction was specific to free will and remained intact after controlling for intelligence (SAT score), Big Five personality traits, and locus of control. Moreover belief in free will was positively correlated with three of the Big Five traits, namely Conscientiousness, Emotional Stability, and Openness to Experience. A field study also measured variations in free will beliefs among mostly poor, low educated, non-white day laborers. Individuals who believed more in free will performed better in these actual jobs, as indicated by ratings by their supervisors (Stillman et al., 2010).

These inter-individual differences in the belief in free will make us wonder if the same variability can be seen in our model comparison and our best-fitting parameter values in Study I. So far, we have analyzed how the inter-individual variability in the reference point parameter can be related to the computed divergence between chosen and unchosen reward. An interesting perspective would be to correlate questionnaires of free will or of locus of control with the best-fitting learning rate values of the counterfactual (CF) model.

7.1.6 A historical perspective

In the introduction, we have showed that the reported locus of control was found to be a stable trait, used in personality psychology to predict people's behavior. Over the past 40 years, locus of control has become one of the most widely studied individual differences in psychology, with most studies using Rotter (1966)'s I-E Scale. In a *Psychology Today* article, Rotter (1971) reported that his samples from the late 1960s and early 1970s were considerably more external than those collected in the early 1960s.

To explore change over time in locus of control, Twenge, Zhang, and Im (2004) examined responses of participants of the same age collected during different years, gathered from the literature. They studied two samples, one of college students and one of children, and found that young Americans increasingly believed their lives were controlled by outside forces rather than their own efforts.

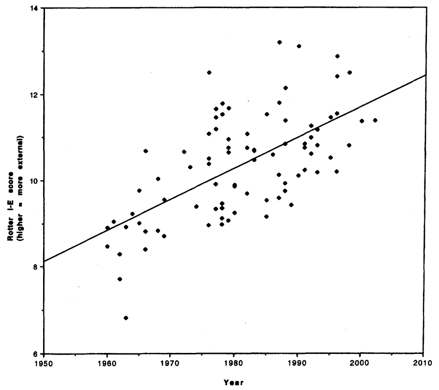


Figure 7.6: College students's locus of control, as measured by the I-E scale, over time. High scores on the scale correspond to a more external locus of control. (Figure reproduced from Twenge, Zhang, and Im, 2004).

The found change was large: the average 2002 student was more external than 80% of college students in the early 1960s, and birth cohort/time period explained 14% of the variance in locus of control. Unfortunately, the implications of increasing externality are almost uniformly negative. A meta-analysis found the self-serving bias to be significantly stronger in individuals with an external locus of control (Campbell and Sedikides, 1999), and this bias is evident in the victim mentality, which was found more common in recent years: Sykes (1992) has thus argued that America has become 'a nation of victims' that blames outrageous behavior on outside sources.

When we interpret our results in Study I, II and III, we should keep in mind that we are studying participants living in a particular culture. As we have seen with this historical perspective on the locus of control, we could imagine that different behavioral trends might have arisen in different periods or from participants from different cultures.

7.2 Formalizing the notion of control

We will now discuss how our results can help formalizing the notion of control. Here we will focus on formalizing people's perception of control, rather than a mathematical or statistical definition of control (see Pearl, 2000 for a review on this subject). We already discussed in the Study I draft how control can be formalized as a difference-making process. The effects of control on behavior and verbal reports have been studied in various areas of psychology, and other mechanisms of control perception have been proposed.

7.2.1 Control as a match between predicted and observed consequences

The notion of control was widely studied in the sensorimotor framework, as an action effectuated via a motor command will always lead to sensory outcomes. Every time our brain sends a motor command, there is evidence that a copy of this command is also generated, called the efference copy (Sperry, 1950). According to a very influential model of sensorimotor control, the predictive forward model, this efference copy will be used to predict the sensory consequences of the action, in order to compare them to the actual perceived consequences. When there is a match between predicted and perceived consequences, a sense of control arises (Frith, Blakemore, and Wolpert, 2000).

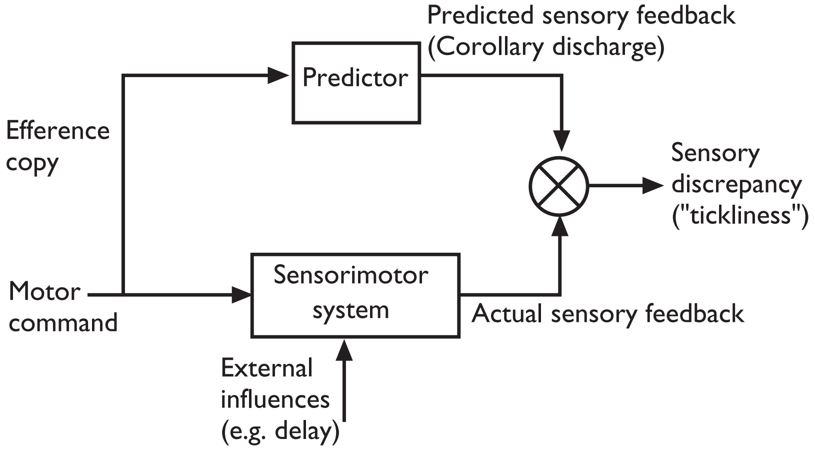


Figure 7.7: A model for determining the sensory consequences of a movement. An internal forward model makes predictions of the sensory feedback based on the motor command. These predictions are then compared with the actual sensory feedback. A mismatch induces a perceived lack of control over the action. (Figure reproduced from Blakemore, Wolpert, and Frith, 2000).

This model was used to explain why you cannot tickle yourself. When a movement is self-produced, its sensory consequences can be accurately predicted, and this prediction can be used to attenuate the sensory effects of the movement. Functional neuroimaging studies have demonstrated that this sensory attenuation might be mediated by somatosensory cortex and anterior cingulate cortex: these areas are activated less by a self-produced tactile stimulus than by the same stimulus when it is externally produced (Blakemore, Wolpert, and Frith, 2000).

Interestingly we have seen in the introduction that the notion of prediction error is also central in the TD(0) algorithm, used to model people's reinforcement learning processes. It was therefore hypothesized, although without providing experimental evidence, that the prediction error as described in the predictive forward model can be linked to the reward prediction error used in a reinforcement learning model (Den Ouden et al., 2008). Therefore one could use the success of the predictive forward model to give support to the 'associative view of causality' that we have developed in Study I, although we rather found evidence for the 'counterfactual view of causality'.

As we said, the formalization of control as a match between predicted and observed consequences was developed in the sensorimotor field, and the efference copy mechanism cannot be generalized to long-term, non sensory outcomes of an action. For example, when one pass an exam, one will feel responsible for the obtained grade, independently of the grade being known one minute or one month after the exam. We would therefore argue that the predictive forward model cannot be generalized outside the sensorimotor framework, and we will now review the other control models that have been developed.

7.2.2 Control as a continuity in the prediction-action-effect chain

According to the predictive forward model, sense of agency arises when external events that follow our action are consistent with predictions of action effects made while we perform or simply intend to perform an action. Thus, agency is inferred retrospectively, after an action has been performed and its consequences are known.

In contrast, a more integrative framework has suggested that internal processes involved in the selection of actions also influence subjective sense of control, in advance of the action itself, and irrespective of effect predictability. Indeed there is evidence that earlier processes, linked to fluency of action selection, prospectively contribute to sense of agency (Chambon, Sidarus, and Haggard, 2014).

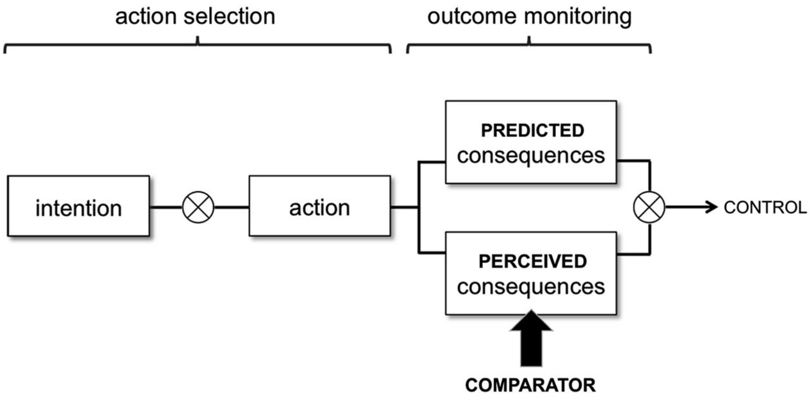


Figure 7.8: The intention-action-effect chain. The action-selection processes operate between the formation of the initial intention and action execution. Dysfluency of action selection signals a break in the intention-action link, and was linked to a decreased perception of control. (Figure reproduced from Chambon, Sidarus, and Haggard, 2014).

More specifically, people feel a stronger sense of control when they choose fluently and easily what to do (Wenke, Fleming, and Haggard, 2010). This result is reminiscent of Lau, Hiemisch, and Baumeister (2015)'s results, mentioned in the introduction: when participants had to choose between three, six or nine housing advertisements, the more options they had, the less free they felt.

At the other end of the spectrum, free-choice, compared to instructed-choice, is known to enhance an induced sense of control. Corah and Boffa (1970) told their subjects that there were two conditions in the experiment, each of which would be signaled by a different light. In one condition they were given a choice of whether or not to press a button to escape from an aversive noise, and in the other one they were not given an opportunity to escape the noise. They found that the choice instructions decreased the aversiveness of the threatening stimulus, apparently by increasing perceived control.

The protocol of Study III was built based on this chain model of control from intention to action. Indeed in Study III, we made the assumption that giving participants the possibility of a choice would enhance their sense of agency, so we could study the link between agency and valence biases in reinforcement learning processes. By forcing participants to match a preselected stimulus, we have thus broken the link between intention and action, and as a consequence we found no choice-confirmatory bias in the participants' learning rates.

A perspective of this work would be to determine whether the selection of a motor action is actually crucial for the choice-confirmation bias to emerge. A future task could be developed in which not pressing a key would lead to the automatic selection of a preselected stimulus. Such protocol could disentangle the importance of action planification and action selection in the choice-confirmation bias. Indeed when a stimulus is preselected and the participant would not press a key, she would still have the intention to choose this stimulus, but without having to generate a motor command.

Our hypothesis would be that in this scenario, not pressing a key would be similar to a choice for participants, and thus a choice-confirmation bias would still appear in these 'passive choice' trials. A parallel can be made with Go/NoGo task, in which inhibiting a Go response is perceived to be a costly and voluntary process (Nieuwenhuis et al., 2003).

7.2.3 Control as instrumental contingency

In the introduction we have seen how control has been defined by the notion of the instrumental contingency(Maier and Seligman, 1976; Hammond, 1980). In most experimental conditions, people's perception of control does correlate with the difference between the probabilities of a consequence knowing the action was performed or not:

$$\Delta p = p(R|A) - p(R|\overline{A}) \tag{7.1}$$

This Δp or contingency model has been very influential to predict animal behavior and human verbal reports (Cheng, 1997), and had inspired many variations of this rule. For example, as the Δp can only be used for binary outcomes, Liljeholm et al. (2013) used a more general metric, the Jensen-Shannon divergence, to compute the difference between probability distributions accounting for more than two possible outcomes.

We also used this notion of contingency when we investigated the effect of a lack of instrumental control on participants' learning strategy in Study II. We implemented a lack of control as a null instrumental contingency ($\Delta p = 0$). Importantly, we explicitly said to participants when they were in control of the action outcomes, and when they were not. We found a significant effect of explicit control on the learning rate parameters, with higher learning rates in the explicit control than lack of control condition (see the additional analyses, Figure 5.4).

Another experiment tested the effect of instrumental control on the learning rate asymmetry (Lefebvre et al., 2016). Although their protocol was slightly different, they also had conditions with control, that they called asymmetric (as the reward probabilities were 25% and 75%) and lack-of-control conditions, called symmetric as the probabilities were the same for both actions (either 25%/25%, or 75%/75%). Their instructions were the same in both conditions, making the lack of control implicit.

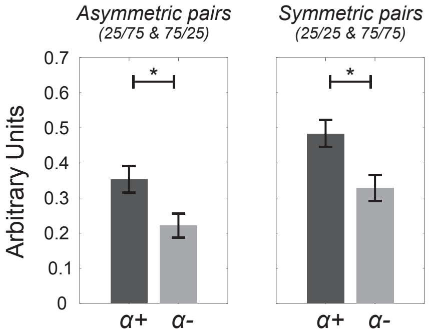


Figure 7.9: Histograms show the learning rates following positive prediction errors (α^+) and negative prediction errors (α^-), obtained from parameters optimization involving only the 'symmetric' (implicit lack of control) or the 'asymmetric' (implicit control) conditions. (Reproduced from Lefebvre et al., 2016, figure S3B)

They also found a significant main effect of condition, but in the opposite direction, as their learning rates were generally higher for 'symmetric' conditions (implicit lack of control) than for 'asymmetric' conditions (implicit control). It is striking to see that explicit and implicit lack of control can have opposite effects, as this is in contradiction with people being able to reliably monitor the implemented contingencies (Liljeholm et al., 2013).

But Matute (1996) has previously shown that humans are able to accurately report their own control only when they are asked to do so at the beginning of the experiment. When participants were not instructed to monitor their action effects (Δp), they actually tended to overestimate their control. Our results in Study I similarly showed that people seemed to rely on a by-default control assumption.

In (Lefebvre et al., 2016)'s implicit lack of control conditions, we interpret participants' behavior as following the local patterns of rewards and adapt their responses to them. Because local patterns are by essence fast-changing, participants must imagine they are in a highly-changing environment to explain the brutal changes in the observed patterns (Yu and Cohen, 2009). They therefore increase their learning rates to adapt to this perceived volatility (Behrens et al., 2007).

But when participants are explicitly told to have no control as in Study II, they don't try to monitor anymore action-outcome associations, thus displaying small learning rates. The by-default control assumption described in Study I can thus explain the differences of effects between implicit and explicit lack of control, found when comparing Study II results to the literature.

7.2.4 Control as cognitive control

Cognitive control, the ability to coordinate thoughts and actions in relation with internal goals, is often required in our everyday life and subserves higher cognition processes such as planning and reasoning (Koechlin, Ody, and Kouneiher, 2003). Cognitive control enables one to inhibit a habitual or automatic response in order to reach a goal in a certain context. For example, imagine you are standing at the corner of a street. Your natural reaction is to look left before crossing, and this is the correct thing to do in most of the world. However, if you are in England, you should repress your 'instinct', and look right. This is a classic example of a circumstance requiring cognitive control (Miller and Cohen, 2001).

Cognitive control and people's perception of control have been mostly studied separately, but our results make us wonder whether a link between the two is possible. Indeed in Study II we found evidence that the post-error boost of implicit agency may be linked to a error-triggered rise in cognitive control. Interestingly in Study I, we found that participants seemed to rely on a by-default control mode. We can thus wonder whether the variations found in the reported sense of agency could be due to the different levels of cognitive control exerted during a task, rather than to the participants' actual monitoring of instrumental control.

A similar link has been made between self-control and the belief in free will, which is closely correlated with the locus of control trait (Baumeister and Brewer, 2012). Believing in free will is apparently tied to a broad sense of wanting to exert control over one's life and believing that one can. That is, believers in free will claim to have better self-control and to be more motivated to exert and maintain control over themselves, as compared to disbelievers in free will.

By contrast, other researchers have focused on the distinction between perceived control either in terms of cognitive control or behavioral control. This parallels the common distinction between cognitive coping ability and behavioral coping ability, frequently applied in the literature on coping with stress (Pearlin and Schooler, 1978). In some situations, people may feel competent to regulate themselves by reappraising the demands or by controlling their emotions, whereas in other situations they may feel competent to change the stressful encounters instrumentally. McCarthy and Newcomb (1992) found that issues such as purpose in life or loss of control were only related to the cognitive control dimension, whereas social stress issues such as assertiveness, leadership, and dating were only related to the behavioral control dimension.

7.2.5 The importance of formalizing control

A fundamental experience of everyday life is the feeling that we control our own actions. When these actions produce effects in the environment, we feel that we cause those too. Contingency awareness, the recognition that components of a future observation can be affected by one's choice of action, is considered a crucial step in the intellectual development of children (Watson and Ramey, 1972). Without this cognitive capacity, it is hard to see how the astonishing range and efficiency of human functional instrumental action could occur. For example, agriculture, material culture and technology all depend on a core cognitive capacity to link one's actions to subsequent effects.

Formalizing the perception of control is an important project not only for psychology, but also for machine learning. Bellemare, Veness, and Bowling (2012) said in a recent article: "While it is not yet clear what mechanisms produce contingency awareness in humans, it seems plausible that some form of contingency awareness could play an important role in the construction of artificially intelligent agents." Indeed, when using popular model-free reinforcement learning algorithms, it is well known that good performance hinges on having access to an appropriate set of basis functions or features (Sutton, 1996). Bellemare, Veness, and Bowling (2012) have proposed a mechanism to identify contingent regions, i.e., the parts of an observation whose immediate future value depends on the the agent's choice. Their results showed that contingency awareness can significantly improve the performance on Atari 2600 games of existing feature construction methods by adding contingency-specific features.

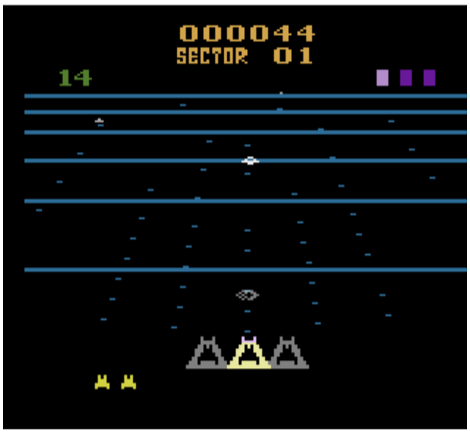


Figure 7.10: Contingency learning by an artificial agent. The contingent regions in Beam Rider (shown in grey) correspond to the avatar's possible next position and missile. (Reproduced from Bellemare, Veness, and Bowling, 2012)

In this PhD thesis, reinforcement learning models were used for the first time to study the relationship between instrumental control, sense of agency and adaptive behavior. Although this work was exploratory, we found this relationship to be rich and diverse, leaving us with new questions to answer. We hope our results can foster future research in this direction.

Remerciements

Je remercie Etienne Koechlin de m'avoir proposé de faire cette thèse, et pour nos nombreux échanges passionants sur le travail de recherche et l'avenir de cette discipline.

Je remercie chaleureusement Valérian Chambon, d'avoir été mon mentor pendant toute ma thèse, de m'avoir appris tnat de choses et d'avoir toujours guidé mes projets.

Je remercie Stefano Palminteri de son soutien pendant toute cette thèse. J'ai énormémemt appris de lui, tant au niveau du design expérimental, de l'analyse des résultats, que du raisonnement scientifique en général.

Je remercie Patrick Haggard de sa passion contagieuse de la science et de m'avoir enseigné en peu de temps tant de choses qui m'auront été utiles tout au long de ma thèse.

Je remercie également toute l'équipe "Action and Body" à UCL, qui ont rendu ma collaboration à Londres très agréable, en particulier Steven di Costa pour son humour et son intégrité scientifique.

Je remercie tous les membres, présents et passés, de notre équipe et du laboratoire d'avoir toujours fourni un feedback de qualité à chacune de mes présentations, avec une mention spéciale à Valentin Wyart qui en donne toujours sans compter.

Cette thèse n'aurait pas pu voir le jour sans le soutien de l'Ecole Doctorale Cerveau, Cognition, Comportement, grâce à laquelle j'ai obtenu ma bourse de thèse, et qui m'a régulièrement encadrée. Je remercie particulièrement Alain Trembleau, Christelle Arruebo et Aude Cortot.

Je remercie mes deux rapporteurs de thèse, Benjamin Scheibehenne et Markus Ullsperger pour le temps qu'ils vont prendre à la relecture de cette thèse, ainsi que Pierre-Yves Oudeyer pour avoir accepté de faire partie de mon jury de thèse.

Je remercie également Mathias Pessiglione et Jérôme Sackur pour leurs conseils lors de mes comités de suivi de thèse, et Benoît Girard pour son soutien.

Je remercie également le bureau des doctorants de l'ENS pour leur aide administrative, en particulier Stéphane Emery.

Je remercie LaTeX pour la mise en page de cette thèse, et au site LaTeXTemplates.com depuis lequel j'ai téléchargé le document-type pour mes manuscrits. Un grand merci à Sci-Hub, sans lequel beaucoup d'articles cités dans cette thèse me seraient restés inconnus. Merci à R et Python, qui ont permis une grande partie des analyses faites dans cette thèse, d'être des logiciels libres. Et bien sûr, merci à

OpenClassRooms de m'avoir appris à utiliser tous ces langages.

Je remercie spécialement mon co-thésard et voisin de bureau, Charles, pour nos pauses thé, nos pauses fruits, nos pauses biscuits et puis nos quelques moments de travail aussi... Merci pour tous ces moments de vie partagés.

Je remercie toutes ces personnes qui font qu'un labo est bien plus que des ordis et des salles de tests : ceux qui aident à chaque fois qu'il leur est possible Gabriel, Margaux, Henri et Aurélien ; mes supers amis du Machine Learning Journal Club Vassilissa Mirjana, Ivan ; et ma fausse stagiaire devenue collaboratrice Sophie. Il serait très ingrat de ma part de ne pas également remercier Marion et Muriel, qui m'ont beaucoup aidée à mon arrivée au labo.

J'aimerais aussi remercier tous les membres du club de Krav Maga de l'ENS, en particulier Olivier Pham, notre Maître. Nos entraînement m'ont permis de libérer intelligemment toutes les tensions qui peuvent survenir dans une thèse. Merci à tous pour votre bonne humeur.

Je remercie ces amis qui rendent la vie plus légère : Pierre-Antoine, Laurène, Charlotte, Alicia, Antonin, Anthony, Marine et Clément. Comme toujours merci à Personne S, pour tout.

Je ne pourrai jamais assez remercier mes parents, mes soeurs et mon frère, de m'avoir appris qu'une famille n'est pas là juste pour les cadeaux à Noël ou aux anniversaires, mais surtout pour s'entraider dans les moments pénibles.

Enfin je remercie Antoine de tout mon coeur, de m'avoir plus appris de choses sur moi-même que toutes la psychologie et les neurosciences ne pourront jamais le faire.

Bibliography

Adams, Henry E and Donald J Lewis (1962). "Electroconvulsive shock, retrograde amnesia, and competing responses." In: *Journal of Comparative and Physiological Psychology* 55.3, p. 299.

Alloy, Lauren B and Lyn Y Abramson (1979). "Judgment of contingency in depressed and nondepressed students: Sadder but wiser?" In: *Journal of experimental psychology: General* 108.4, p. 441.

Alquist, Jessica L et al. (2015). "The making of might-have-beens: Effects of free will belief on counterfactual thinking". In: *Personality and Social Psychology Bulletin* 41.2, pp. 268–283.

Aronson, Elliot (1997). "Back to the future: Retrospective review of Leon Festinger's–A Theory of Cognitive Dissonance". In: *The American Journal of Psychology* 110.1, p. 127.

Bartra, Oscar, Joseph T McGuire, and Joseph W Kable (2013). "The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value". In: *Neuroimage* 76, pp. 412–427.

Baumeister, Roy F and Lauren E Brewer (2012). "Believing versus disbelieving in free will: Correlates and consequences". In: *Social and Personality Psychology Compass* 6.10, pp. 736–745.

Baumeister, Roy F, EJ Masicampo, and C Nathan DeWall (2009). "Prosocial benefits of feeling free: Disbelief in free will increases aggression and reduces helpfulness". In: *Personality and social psychology bulletin* 35.2, pp. 260–268.

Bayes and Price (1763). "An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs". In: *Philosophical Transactions (1683-1775)*, pp. 370–418.

Bayley, Peter J, Jennifer C Frascino, and Larry R Squire (2005). "Robust habit learning in the absence of awareness and independent of the medial temporal lobe". In: *Nature* 436.7050, p. 550.

Behrens, Timothy EJ et al. (2007). "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9, pp. 1214–1221.

Bellemare, Marc G, Joel Veness, and Michael Bowling (2012). "Investigating Contingency Awareness Using Atari 2600 Games." In: *AAAI*.

Berridge, Kent C and Morten L Kringelbach (2008). "Affective neuroscience of pleasure: reward in humans and animals". In: *Psychopharmacology* 199.3, pp. 457–480.

Berry, Dianne C and Donald E Broadbent (1984). "On the relationship between task performance and associated verbalizable knowledge". In: *The Quarterly Journal of Experimental Psychology Section A* 36.2, pp. 209–231.

Bialek, William (2005). “Should you believe that this coin is fair?” In: *arXiv preprint q-bio/0508044*.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blackmore, Susan and Tom Trościanko (1985). “Belief in the paranormal: Probability judgements, illusory control, and the ‘chance baseline shift’”. In: *British Journal of Psychology* 76.4, pp. 459–468.

Blakemore, Sarah-Jayne, Daniel Wolpert, and Chris Frith (2000). “Why can’t you tickle yourself?” In: *Neuroreport* 11.11, R11–R16.

Blanco, Fernando, Helena Matute, and Miguel A Vadillo (2011). “Making the uncontrollable seem controllable: The role of action in the illusion of control”. In: *Quarterly Journal of Experimental Psychology* 64.7, pp. 1290–1304.

Box, George EP (1979). “Robustness in the strategy of scientific model building”. In: *Robustness in statistics*. Elsevier, pp. 201–236.

Brehm, Jack W (1956). “Postdecision changes in the desirability of alternatives.” In: *The Journal of Abnormal and Social Psychology* 52.3, p. 384.

Busemeyer, Jerome R and Adele Diederich (2010). *Cognitive modeling*. Sage.

Campbell, W Keith and Constantine Sedikides (1999). “Self-threat magnifies the self-serving bias: A meta-analytic integration.” In: *Review of general Psychology* 3.1, p. 23.

Cazé, Romain D and Matthijs AA van der Meer (2013). “Adaptive properties of differential learning rates for positive and negative outcomes”. In: *Biological cybernetics* 107.6, pp. 711–719.

Chambon, Valérian, Nura Sidarus, and Patrick Haggard (2014). “From action intentions to action effects: how does the sense of agency come about?” In: *Frontiers in human neuroscience* 8, p. 320.

Chater, Nick and Mike Oaksford (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. OUP Oxford.

Chatlosh, DL, DJ Neunaber, and EA Wasserman (1985). “Response-outcome contingency: Behavioral and judgmental effects of appetitive and aversive outcomes with college students”. In: *Learning and Motivation* 16.1, pp. 1–34.

Cheng, Patricia W (1997). “From covariation to causation: a causal power theory.” In: *Psychological review* 104.2, p. 367.

Clegg, Benjamin A, Gregory J DiGirolamo, and Steven W Keele (1998). “Sequence learning”. In: *Trends in cognitive sciences* 2.8, pp. 275–281.

Colas, Cédric, Olivier Sigaud, and Pierre-Yves Oudeyer (2018). “GEP-PG: Decoupling Exploration and Exploitation in Deep Reinforcement Learning Algorithms”. In: *arXiv preprint arXiv:1802.05054*.

Corah, Norman L and Joseph Boffa (1970). “Perceived control, self-observation, and response to aversive stimulation.” In:

Cox, Richard T (1946). “Probability, frequency and reasonable expectation”. In: *American journal of physics* 14.1, pp. 1–13.

Cross, K Patricia (1977). “Not can, but will college teaching be improved?” In: *New Directions for Higher Education* 1977.17, pp. 1–15.

Dag, Ihsan (1999). “The relationships among paranormal beliefs, locus of control and psychopathology in a Turkish college sample”. In: *Personality and Individual Differences* 26.4, pp. 723–737.

Danielmeier, Claudia and Markus Ullsperger (2011). “Post-error adjustments”. In: *Frontiers in psychology* 2, p. 233.

Daw, Nathaniel D et al. (2006). “Cortical substrates for exploratory decisions in humans”. In: *Nature* 441.7095, p. 876.

Dayan, Peter and Laurence F Abbott (2001). *Theoretical neuroscience*. Vol. 806. Cambridge, MA: MIT Press.

Den Ouden, Hanneke EM et al. (2008). “A dual role for prediction error in associative learning”. In: *Cerebral cortex* 19.5, pp. 1175–1185.

Donoho, David L et al. (2009). “Reproducible research in computational harmonic analysis”. In: *Computing in Science & Engineering* 11.1.

Donoso, Maël, Anne GE Collins, and Etienne Koechlin (2014). “Foundations of human reasoning in the prefrontal cortex”. In: *Science* 344.6191, pp. 1481–1486.

Dudley, R Thomas (1999). “The effect of superstitious belief on performance following an unsolvable problem”. In: *Personality and Individual Differences* 26.6, pp. 1057–1064.

Egan, Louisa C, Paul Bloom, and Laurie R Santos (2010). “Choice-induced preferences in the absence of choice: Evidence from a blind two choice paradigm with young children and capuchin monkeys”. In: *Journal of Experimental Social Psychology* 46.1, pp. 204–207.

Eldar, Eran and Yael Niv (2015). “Interaction between emotional state and learning underlies mood instability”. In: *Nature communications* 6, p. 6149.

Ernst, Marc O and Martin S Banks (2002). “Humans integrate visual and haptic information in a statistically optimal fashion”. In: *Nature* 415.6870, p. 429.

European Commission (2010). “Special Eurobarometer 340: Science and Technology. EBS Report No. 340.” In: *Brussels: European Commission.*

Faraday, Michael (1853). “Experimental investigation of table-moving”. In: *Journal of the Franklin Institute* 56.5, pp. 328–333.

Festinger, Leon (1957). *A theory of cognitive dissonance*. Stanford university press.

Festinger, Leon and James M Carlsmith (1959). “Cognitive consequences of forced compliance.” In: *The journal of abnormal and social psychology* 58.2, p. 203.

Filevich, Elisa et al. (2013). “Brain correlates of subjective freedom of choice”. In: *Consciousness and Cognition* 22.4, pp. 1271–1284.

Fletcher, Paul C et al. (2004). “On the benefits of not trying: Brain activity and connectivity reflecting the interactions of explicit and implicit sequence learning”. In: *Cerebral cortex* 15.7, pp. 1002–1015.

Fouts, Roger and Stephen Tukel Mills (1997). *Next of kin: What chimpanzees have taught me about who we are*. Avon Books.

Frank, Michael J, Lauren C Seeberger, and Randall C O’reilly (2004). “By carrot or by stick: cognitive reinforcement learning in parkinsonism”. In: *Science* 306.5703, pp. 1940–1943.

Frith, Chris (2013). *Making up the mind: How the brain creates our mental world*. John Wiley & Sons.

Frith, Christopher D, Sarah-Jayne Blakemore, and Daniel M Wolpert (2000). “Abnormalities in the awareness and control of action”. In: *Phil. Trans. R. Soc. Lond. B* 355.1404, pp. 1771–1788.

Gaissmaier, Wolfgang et al. (2016). “Betting on illusory patterns: Probability matching in habitual gamblers”. In: *Journal of gambling studies* 32.1, pp. 143–156.

Garrison, Jane, Burak Erdeniz, and John Done (2013). “Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies”. In: *Neuroscience & Biobehavioral Reviews* 37.7, pp. 1297–1310.

Gelman, Andrew et al. (2013). *Bayesian data analysis*. CRC press.

Gershman, Samuel J (2015). “Do learning rates adapt to the distribution of rewards?” In: *Psychonomic bulletin & review* 22.5, pp. 1320–1327.

Gigerenzer, Gerd, Peter M Todd, and the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford University Press.

Gläscher, Jan P and John P O’Doherty (2010). “Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data”. In: *Wiley Interdisciplinary Reviews: Cognitive Science* 1.4, pp. 501–510.

Haggard, Patrick and Valerian Chambon (2012). “Sense of agency”. In: *Current Biology* 22.10, R390–R392.

Haggard, Patrick, Sam Clark, and Jeri Kalogeras (2002). “Voluntary action and conscious awareness”. In: *Nature neuroscience* 5.4, p. 382.

Hammond, Lynn J (1980). “The effect of contingency upon the appetitive conditioning of free-operant behavior”. In: *Journal of the experimental analysis of behavior* 34.3, pp. 297–304.

Heath, Robert G (1972). “Pleasure and brain activity in man”. In: *J Nerv Ment Dis* 154.363, p. 9.

Helmholtz, H von (1856). “Treatise of physiological optics: Concerning the perceptions in general”. In: *Classics in psychology*, pp. 79–127.

Henslin, James M (1967). “Craps and magic”. In: *American Journal of Sociology* 73.3, pp. 316–330.

Holroyd, Clay B and Michael GH Coles (2002). “The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity.” In: *Psychological review* 109.4, p. 679.

Hoorens, Vera (1993). “Self-enhancement and superiority biases in social comparison”. In: *European review of social psychology* 4.1, pp. 113–139.

Hume, David (1739). *A treatise of human nature*. London: Oxford University Press.

Irwin, Harvey J (1992). “Origins and functions of paranormal belief: the role of childhood trauma and interpersonal control.” In: *Journal of the American Society for Psychical Research*.

— (1993). “Belief in the paranormal: A review of the empirical literature”. In: *Journal of the american society for Psychical research* 87.1, pp. 1–39.

Irwin, Harvey J and Caroline A Watt (2007). *An introduction to parapsychology*. McFarland.

Iyengar, Sheena S and Mark R Lepper (2000). “When choice is demotivating: Can one desire too much of a good thing?” In: *Journal of personality and social psychology* 79.6, p. 995.

Jenkins, Herbert M and William C Ward (1965). “Judgment of contingency between responses and outcomes.” In: *Psychological Monographs: General and Applied* 79.1, p. 1.

Jeronimus, BF et al. (2016). "Neuroticism's prospective association with mental disorders halves after adjustment for baseline symptoms and psychiatric history, but the adjusted association hardly decays with time: a meta-analysis on 59 longitudinal/prospective studies with 443 313 participants". In: *Psychological medicine* 46.14, pp. 2883–2906.

Judge, Timothy A (2009). "Core self-evaluations and work success". In: *Current Directions in Psychological Science* 18.1, pp. 58–62.

Keinan, Giora (2002). "The effects of stress and desire for control on superstitious behavior". In: *Personality and Social Psychology Bulletin* 28.1, pp. 102–108.

Knight, David C, Hanh T Nguyen, and Peter A Bandettini (2003). "Expression of conditional fear with and without awareness". In: *Proceedings of the National Academy of Sciences* 100.25, pp. 15280–15283.

Knill, David C and Whitman Richards (1996). *Perception as Bayesian inference*. Cambridge University Press.

Koechlin, Etienne, Chrystele Ody, and Frédérique Kouneiher (2003). "The architecture of cognitive control in the human prefrontal cortex". In: *Science* 302.5648, pp. 1181–1185.

Körding, Konrad P and Daniel M Wolpert (2004). "Bayesian integration in sensorimotor learning". In: *Nature* 427.6971, p. 244.

Kornhuber, Hans-Helmut and L Deecke (1965). "Changes in the brain potential in voluntary movements and passive movements in man: readiness potential and reafferent potentials". In: *Pflugers Archiv fur die gesamte Physiologie des Menschen und der Tiere* 284, pp. 1–17.

Krieghoff, Veronika et al. (2011). "Neural and behavioral correlates of intentional actions". In: *Neuropsychologia* 49.5, pp. 767–776.

Kuchibhotla, Kishore and al. (2018). "Dissociating task acquisition from expression during learning reveals latent knowledge". In: *in prep.*

Langer, Ellen J (1975). "The illusion of control." In: *Journal of personality and social psychology* 32.2, p. 311.

Langer, Ellen J and Judith Rodin (1976). "The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting." In: *Journal of personality and social psychology* 34.2, p. 191.

Langer, Ellen J and Jane Roth (1975). "Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task." In: *Journal of personality and social psychology* 32.6, p. 951.

Laplace, Pierre Simon (1820). *Théorie analytique des probabilités*. Courcier.

Lau, Stephan, Anette Hiemisch, and Roy F Baumeister (2015). "The experience of freedom in decisions–Questioning philosophical beliefs in favor of psychological determinants". In: *Consciousness and cognition* 33, pp. 30–46.

Lefebvre, Germain et al. (2016). "Asymmetric reinforcement learning: computational and neural bases of positive life orientation". In: *bioRxiv*, p. 038778.

Lefebvre, Germain et al. (2017). "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4, p. 0067.

Libet, Benjamin (1985). "Unconscious cerebral initiative and the role of conscious will in voluntary action". In: *Behavioral and brain sciences* 8.4, pp. 529–539.

Libet, Benjamin (1999). "Do we have free will?" In: *Journal of consciousness studies* 6.8-9, pp. 47–57.

Liljeholm, Mimi et al. (2011). "Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction". In: *Journal of Neuroscience* 31.7, pp. 2474–2480.

Liljeholm, Mimi et al. (2013). "Neural correlates of the divergence of instrumental probability distributions". In: *Journal of Neuroscience* 33.30, pp. 12519–12527.

Lindeman, Marjaana and Annika M Svedholm (2012). "What's in a term? Paranormal, superstitious, magical and supernatural beliefs by any other name would mean the same." In: *Review of General Psychology* 16.3, p. 241.

Luu, Long and Alan A Stocker (2018). "Post-decision biases reveal a self-consistency principle in perceptual inference". In: *eLife* 7, e33334.

Lynn, Steven J, Judith W Rhue, and John R Weekes (1990). "Hypnotic involuntariness: A social cognitive analysis." In: *Psychological Review* 97.2, p. 169.

Maier, Steven F (1970). "Failure to escape traumatic electric shock: Incompatible skeletal-motor responses or learned helplessness?" In: *Learning and motivation* 1.2, pp. 157–169.

Maier, Steven F and Martin E Seligman (1976). "Learned helplessness: Theory and evidence." In: *Journal of experimental psychology: general* 105.1, p. 3.

Maier, Steven F and Martin EP Seligman (2016). "Learned helplessness at fifty: Insights from neuroscience." In: *Psychological review* 123.4, p. 349.

Maier, Steven F and Linda R Watkins (1998). "Stressor controllability, anxiety, and serotonin". In: *Cognitive Therapy and Research* 22.6, pp. 595–613.

— (2005). "Stressor controllability and learned helplessness: the roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor". In: *Neuroscience & Biobehavioral Reviews* 29.4, pp. 829–841.

Matute, Helena (1996). "Illusion of control: Detecting response-outcome independence in analytic but not in naturalistic conditions". In: *Psychological Science* 7.5, pp. 289–293.

McCarthy, William J and Michael D Newcomb (1992). "Two dimensions of perceived self-efficacy: Cognitive control and behavioral coping ability". In: *Self-efficacy: Thought control of action*, pp. 39–64.

Medin, Douglas L and Marguerite M Schaffer (1978). "Context theory of classification learning." In: *Psychological review* 85.3, p. 207.

Miller, Earl K and Jonathan D Cohen (2001). "An integrative theory of prefrontal cortex function". In: *Annual review of neuroscience* 24.1, pp. 167–202.

Moore, David W (2005). "Three in four Americans believe in paranormal". In: *Gallup News Service*, p. 161.

Moore, James W and Sukhvinder S Obhi (2012). "Intentional binding and the sense of agency: a review". In: *Consciousness and cognition* 21.1, pp. 546–561.

Morris, John S, Arne Öhman, and Raymond J Dolan (1998). "Conscious and unconscious emotional learning in the human amygdala". In: *Nature* 393.6684, p. 467.

Mueller, Claudia M and Carol S Dweck (1998). "Praise for intelligence can undermine children's motivation and performance." In: *Journal of personality and social psychology* 75.1, p. 33.

Nassar, Matthew R et al. (2010). “An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment”. In: *Journal of Neuroscience* 30.37, pp. 12366–12378.

Ng, Thomas WH, Kelly L Sorensen, and Lillian T Eby (2006). “Locus of control at work: a meta-analysis”. In: *Journal of organizational Behavior* 27.8, pp. 1057–1087.

Nieuwenhuis, Sander et al. (2003). “Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency”. In: *Cognitive, affective, & behavioral neuroscience* 3.1, pp. 17–26.

Nissen, Mary Jo and Peter Bullemer (1987). “Attentional requirements of learning: Evidence from performance measures”. In: *Cognitive psychology* 19.1, pp. 1–32.

Niv, Yael et al. (2012). “Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain”. In: *Journal of Neuroscience* 32.2, pp. 551–562.

O’Doherty, John et al. (2004). “Dissociable roles of ventral and dorsal striatum in instrumental conditioning”. In: *science* 304.5669, pp. 452–454.

Olds, James (1956). “Pleasure centers in the brain”. In: *Scientific American* 195.4, pp. 105–117.

Olds, James and Peter Milner (1954). “Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain.” In: *Journal of comparative and physiological psychology* 47.6, p. 419.

Oudeyer, Pierre-Yves (2018). “Computational Theories of Curiosity-Driven Learning”. In: *arXiv preprint arXiv:1802.10546*.

Overmier, J Bruce and Russell C Leaf (1965). “Effects of discriminative Pavlovian fear conditioning upon previously or subsequently acquired avoidance responding.” In: *Journal of Comparative and Physiological Psychology* 60.2, p. 213.

Overmier, J Bruce and Martin E Seligman (1967). “Effects of inescapable shock upon subsequent escape and avoidance responding.” In: *Journal of comparative and physiological psychology* 63.1, p. 28.

O’Doherty, John P (2014). “The problem with value”. In: *Neuroscience & Biobehavioral Reviews* 43, pp. 259–268.

Palminteri, Stefano, Valentin Wyart, and Etienne Koechlin (2017). “The importance of falsification in computational cognitive modeling”. In: *Trends in cognitive sciences* 21.6, pp. 425–433.

Palminteri, Stefano et al. (2012). “Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning”. In: *Neuron* 76.5, pp. 998–1009.

Palminteri, Stefano et al. (2015). “Contextual modulation of value signals in reward and punishment learning”. In: *Nature communications* 6, p. 8096.

Palminteri, Stefano et al. (2017). “Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing”. In: *PLoS computational biology* 13.8, e1005684.

Paulhus, Delroy L and Jasmine M Carey (2011). “The FAD–Plus: Measuring lay beliefs regarding free will and related constructs”. In: *Journal of personality assessment* 93.1, pp. 96–104.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. Cambridge university press.

Pearlin, Leonard I and Carmi Schooler (1978). “The structure of coping”. In: *Journal of health and social behavior*, pp. 2–21.

Pessiglione, Mathias et al. (2006). "Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans". In: *Nature* 442.7106, p. 1042.

Pessiglione, Mathias et al. (2008). "Subliminal instrumental conditioning demonstrated in the human brain". In: *Neuron* 59.4, pp. 561–567.

Phares, E Jerry (1957). "Expectancy changes in skill and chance situations." In: *The Journal of Abnormal and Social Psychology* 54.3, p. 339.

Portenoy, Russell K et al. (1986). "Compulsive thalamic self-stimulation: a case with metabolic, electrophysiologic and behavioral correlates". In: *Pain* 27.3, pp. 277–290.

Pouget, Alexandre et al. (2013). "Probabilistic brains: knowns and unknowns". In: *Nature neuroscience* 16.9, p. 1170.

Rakos, Richard F et al. (2008). "Belief in free will: Measurement and conceptualization innovations". In: *Behavior and Social Issues* 17.1, p. 20.

Ridderinkhof, K Richard et al. (2004). "The role of the medial frontal cortex in cognitive control". In: *science* 306.5695, pp. 443–447.

Roberts, Seth and Harold Pashler (2000). "How persuasive is a good fit? A comment on theory testing." In: *Psychological review* 107.2, p. 358.

Rodin, Judith and Ellen J Langer (1977). "Long-term effects of a control-relevant intervention with the institutionalized aged." In: *Journal of personality and social psychology* 35.12, p. 897.

Rotter, Julian B (1966). "Generalized expectancies for internal versus external control of reinforcement." In: *Psychological monographs: General and applied* 80.1, p. 1.

Rozin, Paul and Edward B Royzman (2001). "Negativity bias, negativity dominance, and contagion". In: *Personality and social psychology review* 5.4, pp. 296–320.

Rudski, Jeffrey (2004). "The illusion of control, superstitious belief, and optimism". In: *Current Psychology* 22.4, pp. 306–315.

Sartre, Jean-Paul (1956). "Being and Nothingness: An Essay on Phenomenological Ontology". In:

Scheibehenne, Benjamin, Rainer Greifeneder, and Peter M Todd (2010). "Can there ever be too many options? A meta-analytic review of choice overload". In: *Journal of Consumer Research* 37.3, pp. 409–425.

Scheibehenne, Benjamin and Bettina Studer (2014). "A hierarchical Bayesian model of the influence of run length on sequential predictions". In: *Psychonomic bulletin & review* 21.1, pp. 211–217.

Scheibehenne, Benjamin, Andreas Wilke, and Peter M Todd (2011). "Expectations of clumpy resources influence predictions of sequential events". In: *Evolution and Human Behavior* 32.5, pp. 326–333.

Schultz, Wolfram, Peter Dayan, and P Read Montague (1997). "A neural substrate of prediction and reward". In: *Science* 275.5306, pp. 1593–1599.

Seligman, Martin E and Steven F Maier (1967). "Failure to escape traumatic shock." In: *Journal of experimental psychology* 74.1, p. 1.

Shanks, David R and Anthony Dickinson (1991). "Instrumental judgment and performance under variations in action-outcome contingency and contiguity". In: *Memory & Cognition* 19.4, pp. 353–360.

Sharot, Tali and Neil Garrett (2016). "Forming beliefs: Why valence matters". In: *Trends in cognitive sciences* 20.1, pp. 25–33.

Sharot, Tali, Christoph W Korn, and Raymond J Dolan (2011). "How unrealistic optimism is maintained in the face of reality". In: *Nature neuroscience* 14.11, pp. 1475–1479.

Shepperd, James, Wendi Malone, and Kate Sweeny (2008). "Exploring causes of the self-serving bias". In: *Social and Personality Psychology Compass* 2.2, pp. 895–908.

Skinner, Burrhus Frederic (1948). "'Superstition'in the pigeon." In: *Journal of experimental psychology* 38.2, p. 168.

— (1992). "Superstition in the pigeon." In: *Journal of Experimental Psychology: General* 121.3, p. 273.

Sperry, Roger Wolcott (1950). "Neural basis of the spontaneous optokinetic response produced by visual inversion." In: *Journal of comparative and physiological psychology* 43.6, p. 482.

Stillman, Tyler F et al. (2010). "Personal philosophy and personnel achievement: Belief in free will predicts better job performance". In: *Social Psychological and Personality Science* 1.1, pp. 43–50.

Sutton, Richard S (1996). "Generalization in reinforcement learning: Successful examples using sparse coarse coding". In: *Advances in neural information processing systems*, pp. 1038–1044.

Sutton, Richard S and Andrew G Barto (1998). *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge.

— (2017). *Reinforcement learning: an introduction*. the MIT Press Cambridge.

Sykes, Charles J (1992). *A nation of victims: The decay of the American character*. Macmillan.

Takahata, Keisuke et al. (2012). "It's not my fault: postdictive modulation of intentional binding by monetary gains and losses". In: *PloS one* 7.12, e53421.

Taylor, Shelley E and Jonathon D Brown (1988). "Illusion and well-being: a social psychological perspective on mental health." In: *Psychological bulletin* 103.2, p. 193.

Taylor, Shelley E, Heidi A Wayment, and Mary A Collins (1993). "Positive illusions and affect regulation." In:

Thorndike, Edward Lee (1911). *Animal intelligence: Experimental studies*. Macmillan.

Tobacyk, Jerome J, Ed Nagot, and Mark Miller (1988). "Paranormal beliefs and locus of control: A multidimensional examination". In: *Journal of personality assessment* 52.2, pp. 241–246.

Twenge, Jean M, Liqing Zhang, and Charles Im (2004). "It's beyond my control: A cross-temporal meta-analysis of increasing externality in locus of control, 1960-2002". In: *Personality and Social Psychology Review* 8.3, pp. 308–319.

Ullsperger, Markus et al. (2010). "Conscious perception of errors and its relation to the anterior insula". In: *Brain Structure and Function* 214.5-6, pp. 629–643.

Vohs, Kathleen D and Jonathan W Schooler (2008). "The value of believing in free will: Encouraging a belief in determinism increases cheating". In: *Psychological science* 19.1, pp. 49–54.

Vulkan, Nir (2000). "An economist's perspective on probability matching". In: *Journal of economic surveys* 14.1, pp. 101–118.

Wasserman, Edward A, DL Chatlosh, and DJ Neunaber (1983). "Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures". In: *Learning and motivation* 14.4, pp. 406–432.

Watson, John S and Craig T Ramey (1972). "Reactions to response-contingent stimulation in early infancy". In: *Merrill-Palmer Quarterly of Behavior and Development* 18.3, pp. 219–227.

Wegner, Daniel M (2002). *The Illusion of Conscious Will*. The MIT press.

— (2003). "The mind's best trick: how we experience conscious will". In: *Trends in cognitive sciences* 7.2, pp. 65–69.

Wegner, Daniel M and Thalia Wheatley (1999). "Apparent mental causation: Sources of the experience of will." In: *American psychologist* 54.7, p. 480.

Weisberg, Michael (2007). "Who is a Modeler?" In: *The British journal for the philosophy of science* 58.2, pp. 207–233.

Weiss, Jay M (1968). "Effects of coping responses on stress." In: *Journal of comparative and physiological psychology* 65.2, p. 251.

— (1971). "Effects of punishing the coping response (conflict) on stress pathology in rats." In: *Journal of Comparative and Physiological Psychology* 77.1, p. 14.

Weiss, Jay M et al. (1981). "Behavioral depression produced by an uncontrollable stressor: relationship to norepinephrine, dopamine, and serotonin levels in various regions of rat brain". In: *Brain Research Reviews* 3.2, pp. 167–205.

Wenke, Dorit, Stephen M Fleming, and Patrick Haggard (2010). "Subliminal priming of actions influences sense of control over effects of action". In: *Cognition* 115.1, pp. 26–38.

Wessel, Jan R. et al. (2014). "Lesions to the prefrontal performance-monitoring network disrupt neural processing and adaptive behaviors after both errors and novelty." In: *Cortex; a journal devoted to the study of the nervous system and behavior* 50, pp. 45–54.

Whitson, Jennifer A and Adam D Galinsky (2008). "Lacking control increases illusory pattern perception". In: *science* 322.5898, pp. 115–117.

Yu, J Angela and Jonathan D Cohen (2009). "Sequential effects: superstition or rational behavior?" In: *Advances in neural information processing systems*, pp. 1873–1880.

Yudkowsky, Eliezer S. (2003). "An Intuitive Explanation of Bayesian Reasoning". In:

Zhang, Shunan, He Crane Huang, and Angela J Yu (2014). "Sequential effects: A Bayesian analysis of prior bias on reaction time and behavioral choice". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 36. 36.

Résumé

Le sentiment d'agentivité est défini comme le sentiment de contrôler nos actions, et à travers elles, les évènements du monde extérieur. Cet ensemble phénoménologique dépend de notre capacité d'apprendre les contingences entre nos actions et leurs résultats, et un algorithme classique pour modéliser cela vient du domaine de l'apprentissage par renforcement. Dans cette thèse, nous avons utilisé l'approche de modélisation cognitive pour étudier l'interaction entre agentivité et apprentissage par renforcement.

Tout d'abord, les participants réalisant une tâche d'apprentissage par renforcement tendent à avoir plus d'agentivité. Cet effet est logique, étant donné que l'apprentissage par renforcement consiste à associer une action volontaire et sa conséquence. Mais nous avons aussi découvert que l'agentivité influence l'apprentissage de deux manières. Le mode par défaut pour apprendre des contingences action-conséquence est que nos actions ont toujours un pouvoir causal. De plus, simplement choisir une action change l'apprentissage de sa conséquence.

En conclusion, l'agentivité et l'apprentissage par renforcement, deux piliers de la psychologie humaine, sont fortement liés. Contrairement à des ordinateurs, les humains veulent être en contrôle, et faire les bons choix, ce qui biaise notre aquisition d'information.

Mots Clés

Agentivité, Contrôle instrumental, Inférence causale, Prise de décision basée sur des valeurs, Modèles d'apprentissage par renforcement, Modèles bayésien

Abstract

Sense of agency or subjective control can be defined by the feeling that we control our actions, and through them effects in the outside world. This cluster of experiences depend on the ability to learn action-outcome contingencies and a more classical algorithm to model this originates in the field of human reinforcement-learning. In this PhD thesis, we used the cognitive modeling approach to investigate further the interaction between perceived control and reinforcement learning.

First, we saw that participants undergoing a reinforcement-learning task experienced higher agency; this influence of reinforcement learning on agency comes as no surprise, because reinforcement learning relies on linking a voluntary action and its outcome. But our results also suggest that agency influences reinforcement learning in two ways. We found that people learn action-outcome contingencies based on a default assumption: their actions make a difference to the world. Finally, we also found that the mere fact of choosing freely shapes the learning processes following that decision.

Our general conclusion is that agency and reinforcement learning, two fundamental fields of human psychology, are deeply intertwined. Contrary to machines, humans do care about being in control, or about making the right choice, and this results in integrating information in a one-sided way.

Keywords

Agency, Instrumental control, Causal inference, Value-based Decision-Making, Reinforcement Learning models, Bayesian models