Title: Stress-sensitive brain computations of task controllability

Authors: Romain Ligneul[1,2*], Zachary Mainen[1], Verena Ly[2,4†], Roshan Cools[2,3*†]

Affiliations:

[1]Champalimaud Research, Champalimaud Centre for the Unknown, Lisbon, Portugal.

[2]Radboud University, Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Nijmegen, Netherlands.

[3]Radboud University Medical Centre, Department of Psychiatry, Nijmegen, Netherlands.
[4]Department of Clinical Psychology; Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands.

* Correspondence to: romain.ligneul@research.fchampalimaud.org, roshan.cools@fcdonders.ru.nl

† Equal contribution.

Abstract: Estimating environmental controllability enables agents to better predict upcoming events and decide when to engage controlled action selection. How does the human brain estimate environmental controllability? Trial-by-trial analysis of choices, decision times and neural activity in an explore-and-predict task demonstrate that humans solve this problem by comparing the predictions of an "actor" model with those of a reduced "spectator" model of their environment. Neural BOLD responses within striatal and medial prefrontal areas tracked the instantaneous difference in the prediction errors generated by these two statistical learning models. BOLD activity in the posterior cingulate, parietal and prefrontal cortices covaried with changes in estimated controllability. Exposure to inescapable stressors biased controllability estimates downward and increased reliance on the spectator model in an anxiety-dependent fashion. Taken together, these findings provide a mechanistic account of controllability estimation and its distortion in stress-related disorders.

INTRODUCTION

Influential theories suggest that the human brain navigates its environment by building predictive models of the world, which in turn fuel cognitive processes such as directed exploration, goal-directed decisions and forward planning (Bastos et al., 2012; Gläscher et al., 2010; Rao and Ballard, 1999). While these internal models can take diverse mathematical forms, their efficiency always depends on the use of task-relevant and cost-efficient state spaces (Hamid et al., 2019; Kim et al., 2019; Niv, 2019). Most often, these state spaces are actually state-action spaces in which the actions of the agent actively contribute to the prediction of upcoming events. For example, a driver must take into account the movement of their hands in order to predict the future position of their car. By contrast, a passenger worried for their safety should ignore their own hands and instead focus on the hands of the driver in order to anticipate potential hazards.

Determining whether an environment is controllable or not is key to decide to which extent one's actions should influence the prediction process, since only controllable environments afford causal influence over state transitions. Controllable contexts thus prompt the use of "actor" models including one's own actions as predictors, whereas uncontrollable contexts prompt the use of simpler "spectator" models linking past and future states of the environment. By gating the causal influence of action selection, controllability likely plays a central role in the engagement of elaborate action selection mechanisms. Supporting this idea, it is well established that prior exposure to controllable contexts promotes proactive and goal-directed strategies in a variety of cognitive tasks (Maier and Seligman, 2016; Moscarello and Hartley, 2017). Conversely, the lack of perceived control over events, especially stressful ones, constitutes a well-established correlate and a potential predictor of prevalent psychiatric disorders involving an increased influence of reactive and habitual behaviours, such as depression, anxiety,

post-traumatic stress or obsessive-compulsive disorders (Cheng et al., 2013; Diener et al., 2009; Gillan et al., 2014; Hammack et al., 2012; Harrow et al., 2009).

Numerous studies have shown that exposure to uncontrollable stressors can induce a state of learned helplessness characterized by the generalization of passive reactions to subsequent challenges (Maier and Seligman, 2016; Moscarello and Hartley, 2017). Evidence indicates that this maladaptive state largely depends on functional changes within the medial prefrontal cortex (mPFC) and the serotonin system (Amat et al., 2005; Bland et al., 2003; Challis et al., 2013; Kerr et al., 2012; Wood et al., 2015). In humans, a handful of neuroimaging experiments have further suggested that the anterior insula and cingulate cortex contribute to the detrimental effects of uncontrollable stressors (*12*, *13*). Beyond stress induction studies, the sense of being in control of one's own actions and their outcomes is known to modulate hemodynamic responses parietal and prefrontal cortices (Haggard, 2017; Kühn et al., 2013; Spengler et al., 2009) and the right temporoparietal junction (TPJ) was found to track the divergence of action-outcome transitions, a feature of controllable environments (Liljeholm et al., 2013).

Yet, little is known about the algorithms by which the brain estimates dynamically to what degree a task is controllable. A general strategy is to estimate controllability by computing the causal effect the agent's own actions have over the environment. Formally, a task can thus be deemed controllable when the transfer entropy (TE) —a generalization of Granger causality to non-linear and discrete systems— linking state and action time-series is positive (Barnett et al., 2009; Vejmelka and Paluš, 2008). By comparing the entropy of observed states given previous states and actions [H(S'|S,A)] to the entropy of observed states given previous states only [H(S'|S)], this information-theoretic quantity isolates the effective causal influence of actions over state transitions (**Figure 1A, Supplemental Text**). Here, we develop a computational model that tracks a dynamic approximation of TE and use it to shed light on the neural mechanisms supporting the ability to estimate task controllability and adapt behaviour accordingly.

Based on this information-theoretic formalism, we designed an explore-and-predict task that allowed us to manipulate task controllability and assess the resulting changes in terms of subjective controllability and prediction accuracy. This new task was first used in behavioural (n=50) and fMRI (n=32) experiments which aimed at: (a) demonstrating that humans are sensitive to controllability in the absence of reinforcers; (b) establishing the dissociation of spectator and actor models at the behavioural and neural levels; and (c) unraveling the neural substrates underlying the representation of controllability itself and its influences on behaviour. In a subsequent stress experiment (n=54), we exposed participants to either uncontrollable or controllable electric shocks before administering the explore-and-predict task in order to (d) provide causal evidence supporting a dissociation of the spectator and actor models and (e) test whether learned helplessness can be characterized by an increased reliance on the former relative to the latter.

RESULTS

Experimental paradigm and computational model

Healthy human participants were invited to explore an abstract environment composed of three states (square, circle, triangle) and three actions (yellow, blue, magenta). A hidden transition rule always determined upcoming states, either dependent on the action of the participants (controllable rules, C) or the previous state only (uncontrollable rules, U) (**Figure 1B**). The transition rules were probabilistic and reversed covertly, so participants needed to explore and accumulate evidence in order to tell which rule was operative. From time to time, the participant was asked to predict the most likely upcoming state given a state-action pair (e.g. “blue” action in “circle” state), and its counterfactual (e.g. “yellow” action in “circle” state). This procedure yielded a direct yet implicit assessment of their subjective sense of controllability, because counterfactual predictions should only differ in controllable contexts, where selected actions determine upcoming states (**Figure 1C**). A novel and distinguishing feature of our task is that controllability varied independently of uncertainty (**Figure 1D**), a methodological improvement over earlier paradigms where the two constructs covary systematically (Diener et al., 2009;

Dorfman and Gershman, 2019; Liljeholm et al., 2013). Another key difference from previous studies is that we did not include any reinforcers: participants were merely instructed to explore their environment with the goal of performing accurate predictions when asked to. Here, controllability estimation can interact with, but does not depend on reward and punishment processing (Huys and Dayan, 2009; Ly et al., 2019; Moscarello and Hartley, 2017), the only requirement being to maintain a minimal level of exploration, or noise in the action selection process (**Figure 1E**, **Supplemental Text**).

To untangle the mechanisms of controllability estimation in this task, we designed a computational architecture for dynamically tracking an approximation of TE (for a detailed description, see **Figure S1A**, **Methods**). Paralleling the standard computation of TE, two sets of transition probabilities were monitored, one corresponding to an "actor" model (tracking state-action-state transition, SAS') and the other to a "spectator" model (tracking state-state transition, SS', **Figure 1F,i**). Following each transition, an approximation of TE (hereafter termed Ω) was updated in proportion to the difference between 'actor' and 'spectator' transition probabilities $p_{SAS'}$-$p_{SS'}$ (**Figure 1F,ii**). Intuitively, this difference term can be understood as an instantaneous causality signal, reflecting how likely the last state transition towards S' was due to the influence of action A rather than state S. By integrating $p_{SAS'}$-$p_{SS'}$ over time, Ω thus reflects the causal influence of actions on recent state transitions (**Figure S1B**).

This causality signal Ω is at the core of the proposed algorithm, which arbitrates between the actor and the spectator model when making predictions about upcoming states. Specifically, the relative weight of the actor versus spectator model is set by an arbitrator (hereafter termed ω) whose value can be interpreted as an estimate of controllability. Two parameters influence the mapping between Ω and ω: a threshold determining how much causal evidence is required to infer controllability and a slope determining how fast controllability estimates change around that threshold (**Figure 1F,iii**). This SAS'-SS'-Ω algorithm was contrasted with a conventional model-based reinforcement learning algorithm monitoring SAS' transitions, as implemented by the actor model alone (Gläscher et al., 2010; Lee et al., 2014). Importantly, this simpler algorithm could still learn

transition probabilities from both uncontrollable and controllable conditions in stable environments, but the lack of controllability-dependent arbitration makes it less efficient in volatile environments alternating rapidly between controllable and uncontrollable rules.

Flexible adjustment of behaviour to changes in controllability

Participants performed well on the task: in all experiments, the average prediction accuracy was substantially above chance [chance level: 1/3; behavioural: $t(49)=12.3$, $p<0.001$; fMRI: $t(31)=13.4$, $p<0.001$; stress: $t(53)=13.2$, $p<0.001$](**Figure 2A**). In the fMRI experiment, for which participants received more training, accuracy was also stable across conditions and time (**Table S1**). Prediction accuracy dropped and then rapidly recovered after covert reversals in transition rules: it already exceeded chance levels on the first prediction trial after reversal for all types of reversals (all $t(49)>5.38$ and all $p<0.001$](**Figure 2B**). Prediction accuracy also correlated positively with working memory capacity as indexed by d-primes in a standard 2-back task [U rules: $\rho=0.37$, $p=0.006$; C rules: $\rho=0.52$, $p<0.001$](**Figure 2C**), consistent with the engagement of a model-based learning process (Otto et al., 2013).

In line with our prediction that humans solve the task by estimating Ω, Bayesian model comparisons demonstrated that SAS'-SS'-Ω schemes outperformed the conventional model-based learning algorithm (SAS' alone) in all experiments (**Figure 3A**; **Figures S2A** and **3**), Simulation analyses confirmed that the model was identifiable and that its parameters could be recovered accurately (**Figures S2B-D**). As expected, the arbitrator ω captured quantitative changes in subjective controllability, indexed by the proneness of participants to predict that different actions would lead to different states in counterfactual prediction trials (**Figure 3B**). Critically, the SAS'-SS'-Ω scheme which included an arbitration mechanism accounted better for the dynamics of subjective controllability changes around reversals than did the SAS' model alone (**Figure 3C**, correct prediction of controllability: 72.7% versus 66.5%, $z(49)=4.69$, $p<0.001$). The benefits of monitoring controllability is further illustrated by the finding that the likelihood of using the SAS'-SS'-

Ω scheme over the SAS scheme increased with accuracy across subjects [behaviour: r=0.58, p<0.001; fMRI: r=0.32, p=0.07; stress: r=0.54, p<0.001] (**Figure S2E-G**).

Behavioural and neural dissociation of the actor and spectator models

Model comparison results are consistent with our proposal that subjects estimate the subjective controllability of an environment by separately tracking and comparing an actor and a spectator model. In order to further test the dissociation of the actor and spectator models, we used subject-level GLMs to assess trial-by-trial fluctuations of decision times. It is known that decision times slow down following state prediction errors (Shahar et al., 2019b, 2019a). The large amount of exploratory trials per participant thus allowed us to analyze decision times as a proxy of model updating and to evaluate to which extent controllability *per se* influences the speed of action selection [behaviour: 562+/-163 trials; fmri: 550+/-115; stress: 519+/-84]. We therefore extracted the prediction errors derived from both the actor and spectator models (hereafter termed $\delta_{SAS'}$ and $\delta_{SS'}$). We found that both type of prediction errors slowed responding [behavioural: betas $\delta_{SAS'}$: t(49)=3.53, p<0.001; betas $\delta_{SS'}$: t(49)=7.93, p<0.001; fMRI: betas $\delta_{SAS'}$: t(31)=3.50, p=0.001; betas $\delta_{SS'}$: t(31)=2.46, p=0.020] (**Figure 3E**) and independently explained variance in decision times (**Figure S4**). We also observed that in periods of higher estimated environmental controllability (i.e. higher ω), decision times were slower [behavioural: t(49)=2.20, p=0.032; fMRI: t(31)=3.78, p<0.001]. This effect suggests that controllable contexts promote a more controlled action selection process even when no reinforcement is at stake.

Separable neural correlates should therefore exist for the prediction errors generated by the actor and spectator probability tracking processes, $\delta_{SAS'}$ and $\delta_{SS'}$. A conjunction analysis first revealed that both types of prediction errors activated the typical set of bilateral brain areas commonly associated with state prediction errors (Gläscher et al., 2010; Lee et al., 2014), such as the frontoparietal network and the pre-supplementary motor area (**Figure 4A**, **Table S2**). When contrasting events where only $\delta_{SS'}$ or only $\delta_{SAS'}$ were above their median value (a contrast circumventing collinearity issues), the mPFC

significantly dissociated the two prediction error terms (**Figure 4B**, **Table S3**, see also **Figures S5A-B** for robustness checks). Mixed-effects ROI analyses indicated that the mPFC encoded $\delta_{SS'}$ more negatively than $\delta_{SAS'}$.

Testing directly the effect of $\delta_{SS'}$-$\delta_{SAS'}$ (mathematically equivalent to $p_{SAS'}$-$p_{SS'}$) using a conventional parametric analysis at the whole brain level showed that the mPFC also negatively encoded this signal required for the update of controllability. Like the mPFC, the nucleus accumbens encoded $\delta_{SS'}$-$\delta_{SAS'}$ negatively, but ROI analyses indicated that this effect stemmed from a positive response to $\delta_{SAS'}$ and an absence of relationship with $\delta_{SS'}$ (**Figure 4C**, **Table S3**). Interestingly, a similar pattern was observed in the dopaminergic nuclei of the brainstem at a more lenient threshold (**Figure S5C**).

Neural correlates of dynamic controllability

Having established the dissociation of $\delta_{SAS'}$ and $\delta_{SS'}$ at the behavioural and neural levels, we next probed the correlates of the prediction error δ_{Ω} governing changes in estimated controllability ($\delta_{\Omega}=\delta_{SS'}-\delta_{SAS'}-\Omega_{t-1}$). This second order learning mechanism is key to accumulate, over time, evidence in favor or against the controllability of the ongoing rule. Whole-brain analyses revealed a significant negative relationship between δ_{Ω} and neurovascular responses in the posterior cingulate (PCC, BA 29/30) and dorsal posterior cingulate cortex (dPCC, BA 23/24), the right dorsal anterior insula (dAI) and the right temporo-parietal junction (TPJ, **Figure 4D**, **Table S4**). Mixed-effects ROI analyses including decision times and Ω confirmed that these effects reflected a genuine response to controllability prediction error, peaking 4-8 seconds after trial onset. Interestingly, we observed a controllability-dependent coupling between the PCC and the mPFC cluster found to encode the difference term $\delta_{SS'}$-$\delta_{SAS'}$ (**Figure S5D**). Moreover, interindividual variability in the proneness to rely on the actor model and to perceive the environment as controllable was partly explained by the strength of δ_{Ω} encoding in this structure [r=-0.47, p=0.007](**Figure 4E, Figure S5E-F**).

In order to unravel the neural correlates of controllability with maximal sensitivity, we performed a multivoxel pattern analysis (MVPA). A support vector machine classifier was trained at predicting whether streaks of consecutive exploratory trials were governed by controllable or uncontrollable rules. Whole-brain maps of classification accuracy were obtained using the searchlight method (leave-one-run-out cross-validation). Objective rule controllability could be decoded above chance from most regions of the frontoparietal control network. Local patterns of activity in the precuneus, the right TPJ, the dlPFC (bilaterally) and the dorsomedial prefrontal cortex (dmPFC) were all sensitive to controllability (**Figure 4F, Table S5**). Interestingly, the sensitivity of the dmPFC to controllability predicted to which extent controllable contexts lengthened decision times from one participant to another [r=0.53, p=0.002](**Figure 4G**).

Uncontrollable stressors promote reliance on the spectator model

We applied this new paradigm to better understand the computational mechanisms underlying learned helplessness. More precisely, we hypothesized that exposure to uncontrollable stressors might bias controllability estimation mechanisms to promote reliance on the spectator model relative to the actor model. We invited participants to perform an active avoidance task exposing them to mild electric shocks prior to completing the explore-and-predict task (**Figure 5A**). Participants in the controllable group learned to avoid the shock following one of the three possible cues by pressing the correct response button (out of six alternatives). Shocks received by participants in the uncontrollable group were yoked to the former, so that their decisions had no influence on shock probability. As expected, this procedure induced a dissociation between actual shock frequency, matched across groups by design, and reported shock expectancy (**Figure 5B**), so that shock expectancy remained high until the end of the induction phase in the uncontrollable group.

Despite the absence of aversive reinforcers in the explore-and-predict task, we observed clear carry-over effects from the shock experiment when analyzing the model parameters governing controllability estimation (**Figure 5C**, **Table S6**). In particular, the threshold

parameter increased significantly in the uncontrollable group compared with the controllable group [t(52)=2.82, p=0.007]. This parameter determines how much causal evidence is required before controllability is inferred. Therefore, when making predictions, participants exposed to uncontrollable stressors relied more on the spectator model, demonstrated by the reduced average value of the arbitrator ω [z=-2.44, p=0.015] as well as the direct analysis of subjective controllability estimates, revealing that counterfactual predictions were more often identical in the uncontrollable group (z=1.69, p=0.045, one-tailed). Importantly, counterfactual predictions did not differ during the training phase, which occurred before stress induction [z=-0.58, p=0.56].

Exposure to uncontrollable stressors thus elicits an imbalance between actor and spectator mechanisms for transition probability learning consistent with a sustained shift in controllability expectations. This finding provides a parsimonious account of the cross-contextual generalization of passive strategies, a core feature of helpless states (Maier and Seligman, 2016; Moscarello and Hartley, 2017). Interestingly, state anxiety, as assessed before the experiment, moderated the induction of controllability estimation biases. It predicted the average value of the arbitrator only in participants exposed to uncontrollable stressors [U: r=-0.47, p=0.014; C: r=0.24, p=0.21, U versus C: z=-2.59, p=0.01](**Figure 5D**).

Since uncontrollable stressors promoted increased reliance on the spectator model and decreased reliance on the actor model, we expected PE-dependent slowing effects to follow a similar pattern. Confirming this prediction, the type of stress induction profoundly altered the slowing of decision times by actor and spectator prediction errors [interaction group by PE type: F(1,52)=9.34]. In the uncontrollable group, $\delta_{SAS'}$ no longer modulated decision times whereas the impact of $\delta_{SS'}$ was increased compared with the controllable group [$\delta_{SS'}$: t(52)=2.08, p=0.042; $\delta_{SAS'}$: t(52)=-3.37,p=0.001]. Similarly, Ω no longer predicted decision times [t(26)=1.04, p=0.15], contrary to what was observed in the controllable group [t(26)=1.98,p=0.029](**Figure 5E**) and in the other two experiments.

DISCUSSION

Taken together, these findings shed light on one of the most fundamental aspects of human experience: the ability to estimate to which extent our actions affect our environment and to adjust our decisions accordingly. Our results demonstrate that this ability involves the comparison of actor and spectator models of the ongoing task, which are dissociable computationally, behaviourally and neurally. In turn, controllability estimates can be used to arbitrate between these models when making predictions about future events. The mPFC and the striatum encode the difference between the prediction errors generated by each model, while signals related to the update of controllability estimates are found in a more posterior brain network encompassing the TPJ and the PCC. Furthermore, exposure to uncontrollable stressors biases this process assessed by the explore-and-predict task, hence establishing its relevance for the study of neuropsychiatric disorders involving altered perceptions of controllability (Cheng et al., 2013; Diener et al., 2009; Gillan et al., 2014; Voss et al., 2017).

Historically, the concept of task controllability has been heavily influenced by learned helplessness studies in which animals granted the ability to actively terminate stressors are compared to yoked animals exposed to the exact sequence of stressors, but whose actions are made independent from stressor termination (Amat et al., 2005; Bland et al., 2003; Maier and Seligman, 2016). In this line of research, focused on the long-lasting consequences of stress exposure, more controllable contexts were defined as those in which the mutual information linking the timings of actions and stressor offsets is higher (Maier and Seligman, 1976). However, a positive mutual information linking an organism's actions and upcoming states of the environment is a necessary but not a sufficient condition to declare a task controllable. For example, the highly positive mutual information linking the statements of a weather forecaster with the occurrence of rain should obviously not be interpreted as a sign that the forecaster controls the weather, because the statements of the forecaster and the occurrence of rain are both conditioned by past meteorological states. Moreover, following this incomplete definition, variations of task controllability were often obtained by manipulating uncertainty about future states

(Bräscher et al., 2016; Dorfman and Gershman, 2019; Kerr et al., 2012), hence leading to ambiguous conclusions regarding the mechanisms underlying the estimation of controllability *per se* and its downstream influence on behaviour.

Formalizing controllability using transfer entropy (TE) rather than mutual information allowed us to design a task in which controllability varied independently from uncertainty. In addition, this approach provided an algorithm for detecting genuine changes in task controllability. Model comparisons showed that, across the three experiments, algorithms monitoring controllability using an approximation of TE (i.e SAS'-SS'-Ω schemes) accounted better for participants' choices than a standard model-based learning algorithm. The analysis of reaction times confirmed that participants were sensitive to the prediction errors generated by the spectator and actor models. These two first-order models, whose comparison governed the update of controllability estimates, can be viewed as two state spaces competing to structure the learning of statistical contingencies. When controllability estimates are low, the spectator model representing only the successive states of the environment dominates. In contrast, when controllability estimates are high, the actor model representing both states and actions takes over.

By defining the most appropriate state space dynamically, controllability estimation improves predictions about the future states of one's environment and can therefore contribute to maximize utility when reward or punishment rates depend on such predictions. By promoting reliance on a simpler spectator model when the environment is deemed uncontrollable, it can also minimize the metabolic cost and subjective effort associated with controlled action selection (Hahn et al., 2020; Westbrook et al., 2020). These hypotheses could be tested directly by introducing reinforcers in the explore-and-predict task, but it is already worth to note that the controllability-dependent arbitration logics can readily explain why Pavlovian (equivalent to SS') and instrumental (equivalent to SAS') learning mechanisms are respectively promoted uncontrollable and controllable contexts (Dorfman and Gershman, 2019). Restricted to the striatum and the mPFC, the limited spatial dissociation of the actor and spectator models reported here is consistent with a recent MEG study showing that the human brain relies on shared "neural codes"

to infer hidden states in controllable and uncontrollable contexts (Weiss et al., 2019). Yet, signatures of the actor and spectator models may ultimately be found within local neural circuits, which are beyond the reach of standard fMRI and MEG techniques. Moreover, the preferential encoding of actor prediction errors by the striatum and dopaminergic midbrain is consistent with earlier findings showing that the mesolimbic pathway preferentially encodes reward prediction errors in instrumental learning tasks (Garrison et al., 2013; Grogan et al., 2020; Hamid et al., 2019; O'Doherty, 2004). Conversely, the preferential encoding of spectator prediction errors by the mPFC is consistent with its more general role in statistical learning (Gilboa and Marlatte, 2017; Klein-Flügge et al., 2019). It may explain why dysfunctions of this structure impair observational learning more profoundly than instrumental learning (Jurado-Parras et al., 2012; Kumaran et al., 2015) and why mPFC lesions can alter the perception of controllability without altering performance of simple instrumental learning tasks (O'Callaghan et al., 2019).

By comparing the predictions emanating from the actor and spectator models, one can derive an instantaneous causality signal (i.e how likely did the last action cause the last state transition). Encoded negatively by mPFC and striatal BOLD responses, this instantaneous signal can then be integrated over time, hence reflecting the causal influence of actions over recent transitions. A signature of the second-order prediction errors supporting this integration was found in the right TPJ, right insula, PCC and dPCC. The right TPJ was the only region sensitive to both these second-order prediction errors as well as to objective controllability as assessed by the decoding analysis. It is thus a strong candidate for the implementation of controllability monitoring in our task. Supporting this view, the right —but not left— TPJ has previously been found to encode the divergence in action-outcome distributions (Liljeholm et al., 2013) and the discrepancy between expected and actual outcome timings in a simple sensorimotor task alternating controllable and uncontrollable trials (Spengler et al., 2009). Other brain areas including the dlPFC and the dmPFC were sensitive to objective controllability, likely reflecting downstream adaptations of brain networks to task controllability (Wanke and Schwabe, 2019). Strikingly, a higher sensitivity of the dmPFC to controllability was associated with a stronger influence of controllability on decision times. This finding suggests that

controllability detection fosters more elaborate action selection processes by potentiating a form of proactive response inhibition previously linked to dmPFC activity (Albares et al., 2014). It might be noted that the dmPFC region discussed here is more dorsal than the dACC region previously implicated in the valuation of control (Shenhav et al., 2016).

Interestingly, the PCC exhibited a controllability-dependent coupling with the mPFC and participants with a more (negative) encoding of controllability prediction errors in the PCC had a higher propensity to rely on the actor model when making predictions. Based on this finding, we may speculate that deactivations of the PCC facilitate the switch towards the actor model whenever the environment is deemed controllable. Supporting this view, electrical stimulation of the human PCC can elicit behavioural idleness and loss of control feelings (Vesuna et al., 2020), a phenomenon possibly due to the sudden blunting of action monitoring mechanisms(Li et al., 2019). Future studies may test whether the heightened metabolic activity of the pCC (Leech and Sharp, 2014) and its unstable connectivity with the mPFC during major depression (Wise et al., 2017) contributes to the lower sense of control observed in this condition (Cheng et al., 2013).

Having described the computational principles and outlined neural correlates of controllability estimation, we sought to test whether an experimental manipulation could alter this process and simultaneously contribute to a better understanding of the learned helplessness phenomenon. Indeed, exposure to uncontrollable stressors is known to induce passive responses to subsequent controllable stressors, but the origins of this maladaptive strategy remain poorly understood. In particular, it is unclear whether prior exposure to uncontrollable stressors induces an increased sensitivity to future aversive events, reduces expectation of control with respect to future stressors or reduces expectations of control in general (Ly et al., 2019; Maier and Seligman, 2016). Our results support the latter hypothesis by showing sustained alterations of controllability estimation in human participants previously exposed to uncontrollable versus controllable stressors. More precisely, the specific increase observed for the threshold parameter implies that the former group needed to integrate more causal evidence before considering a given rule as controllable in the explore-and-predict task. The dorsal anterior insula (dAI) is

involved in the modulation of pain perception by controllability (Bräscher et al., 2016) and it was found to encode controllability prediction errors in our fMRI experiment. Therefore, it is possible that prior exposure to uncontrollable stressors altered dAI excitability to distort subsequent controllability estimation mechanisms. Supporting this idea, a study showed that a lower perception of control mediates the exacerbation of dAI responses to physical threats in more anxious individuals, who also displayed lower controllability estimates following uncontrollable stressors in our data (Alvarez et al., 2015). Yet, this increased reliance on the spectator relative to the actor model following uncontrollable stressors likely involves several other brain areas, including the mPFC and the dorsal raphe nucleus, both highly sensitive to stressor controllability (Amat et al., 2005; Bland et al., 2003; Challis et al., 2013).

In sum, the explore-and-predict task allowed us to isolate the core computations underlying controllability estimation by excluding reinforcers and matching uncertainty across contexts. By showing that the human brain can compute an approximation of transfer entropy, our study may help bridging the gap between neuroscience and artificial intelligence research, where transfer entropy plays an important role in solving unsupervised learning problems (Leibfried et al., 2020; Mohamed and Jimenez Rezende, 2015). More invasive techniques will be required to understand how these computations are implemented within local neural circuits and how neuromodulators such as dopamine or serotonin mediate their broad impact on stress responses and mental health (Amat et al., 2005; Bland et al., 2003; Dickerson and Kemeny, 2004; Moscarello and Hartley, 2017).

STAR Methods

Participants

For the behavioural experiment, fifty young adult participants (mean age: 24.7, range: 18—43, 27 women) were recruited via the Sona system (human subject pool management system) of the Radboud University (The Netherlands). All participants were included in the data analysis. For the fMRI experiment, thirty-two young adult participants (mean age: 25.1, range: 20—43, 18 women) were recruited through the same system. For the stress experiment, a total of 62 participants (mean age = 21.8; range: 18-27, 52 women) were recruited via the Sona system of Leiden University. One additional participant was excluded a posteriori from the fMRI experiment and four participants were excluded from the stress experiments, together with their yoked counterparts (see Supplemental Methods for details on exclusion and inclusion criteria). The behavioural and fMRI experiments were approved by the local ethics committee (CMO region Arnhem/Nijmegen, The Netherlands, CMO2001/095). The stress experiment was approved by the Psychology Research Ethics Committee (CEP17-0905/282) at Leiden University. All participants provided written informed consent, in line with the declaration of Helsinki.

Explore-and-predict task

In the 3 experiments, the overall structure of the task was identical. Participants performed 6 (fMRI and stress experiment) or 7 (behavioural experiment) exploratory trials before a pair of predictions was required. Pairs of predictions always probed the two actions available for a given state (e.g blue followed by yellow in the circle state), in order to derive subjective control lability from counterfactual responses. Participants received feedback about their predictions in 50 percent (fMRI and stress) or 100 percent (behavioural experiment) of the trails. In the fMRI and stress tasks, feedback was delivered only after one of the two counterfactual predictions in order to prevent participants from inferring whether the rule was controllable or not based on feedback.

On each exploratory trial, two identical geometrical shapes were displayed side by side. The color of each shape determined the action corresponding to left and right button presses (side randomly assigned in each trial). An urgency signal was displayed after 1.5s. Transitions to the next state were always governed by one of the four rules, as displayed in Figure 1C. In order to maximize the variation of prediction errors, the transitions were stochastic (noise: 0.05 to 0.2).

The first prediction trial of each pair was simply displayed at the end of the ITI of the previous exploratory trial. An urgency signal was displayed after 4s. The hypothetical state action pair was displayed at the center of the screen, just below a question mark, and the 3 possible next states were displayed as white geometrical shapes at the top of the screen. The selected state was then highlighted and the feedback was displayed when applicable.

The ongoing rule was never changed before 4 pairs of predictions were completed. In the behavioural experiment, the rule changed from then as soon as 5 correct responses were provided in the last 6 predictions or if the last 4 predictions were accurate. In the fMRI experiment, the rule was changed as soon as the p-value of a binomial test indicated that accuracy was significantly below chance ($p<0.05$, one-tailed, chance level: 1/3), hence making the accuracy threshold more lenient as the number of predictions made for a given rule increased. In all experiments, the rule changed after 10 pairs of predictions, even if performance did not meet the learning criterion. Prediction trials were pseudo-randomly ordered with the constraint that each state would be tested a similar number of times. An exhaustive description of instructions, counterbalancing, reversal schedules, transition noises and timings is available in Supplemental Information.

Stress induction task

To test the impact of prior controllability over stress on subsequent controllability estimations, participants underwent a stressor controllability manipulation prior to the

explore-and prediction task in a between-subjects design. Critically, we employed a between-subjects yoked control procedure in order to match the amount and order of aversive outcome stimuli between the controllable and uncontrollable conditions. We randomized participants in blocks of four where the controllable condition of a yoked pair was always administered first in order to create the schedule for the yoked counterpart in the uncontrollable condition.

Electric stimuli served as stressors in the manipulation task and were delivered by a Digitimer DS7 stimulator. First, individual levels of intensity of the electric stimulus for the manipulation task were determined using a stepwise procedure in which the intensity of the stimulus was gradually increased until participants reported a 'just bearable, but not yet painful' experience of shock. A yoked control-design with preprogrammed pseudorandomized schedule enabled us to match the amount and order of electric stimuli between the conditions.

In the controllable condition, a total of four cues (different in shape and color) were presented for at least six repetitions each following a preprogrammed pseudorandomized schedule. Participants could learn by trial-and-error the correct response corresponding to the cue (a key between 1 and 6) to avoid the electric stimulus. Critical trials on which participants would be able to prevent the electric stimulus for the first time according to the schedule were repeated until the participants arrived at a correct response. As such, all participants underwent the whole schedule with a minimum of 24 trials, and were able to acquire the correct response for each cue.

The uncontrollable condition was yoked to the controllable condition, such that participants experienced a comparable pattern of events across conditions. However, in the uncontrollable condition, participants were not able to acquire these action-outcome contingencies to prevent the shocks, whose sequences were merely replayed from the yoked participants performing the controllable condition. An exhaustive description of counterbalancing, instructions and procedures is available in Supplemental Information.

Computational modeling

The main purpose of all SAS'-SS'-Ω algorithms is to provide a way to dynamically estimate the causal influence of actions over state transitions by updating a variable termed Ω. In all models, S represents the previous state of the environment, A represents the previous action and S' represents the current state of the environment. The local causality estimate Ω can only be used as a proxy for controllability, which is not a property of actions but of the environment. It is this "inferred controllability" variable, termed ω, which can then be used to decide (arbitrate) whether one should make predictions using learned S-S' transitions or learned SA-S' transitions. Ω is homologous to transfer entropy (TE, which is itself a generalization of Granger causality to discrete and non-linear domains). See Supplemental Methods for a detailed explanation of the differences between TE and Ω.

In order to demonstrate that participants used a dynamic estimate of transfer entropy to solve the task, we systematically compared variants of the SAS'-SS'-Ω algorithm to a standard model based architecture tracking SAS' transitions (Gläscher et al., 2010). This latter algorithm corresponds to the actor model alone. Its asymptotic performance in stable environments is identical to that of SAS'-SS'-Ω algorithms.

The actor model tracks transitions linking state-action pairs to newly encountered states (i.e. SAS'). It updates transition probabilities in the following fashion.
Realized transitions:

$$P(s'|s,a) \leftarrow P(s'|s,a) + \alpha(1 - P(s'|s,a))$$

Unrealized transitions:

$$P(s'|s,a) \leftarrow P(s'|s,a)(1 - \alpha)$$

Where $\alpha \in [0,1]$ controls to which extent learned transition probabilities are determined by the most recent transitions. The prediction error $1-P(s'|\, s,a)$ is noted $\delta_{SAS'}$ in the main text.

The spectator model tracks transitions linking states to newly encountered states (i.e. SS'). Therefore, it updates transition probabilities exactly like the actor model, except that only states are represented: P(s'| s,a) is simply replaced by P(s'| s) in the two equations above, and the prediction error 1−P(s'| s) is noted $\delta_{SS'}$ in the main text.

The variable Ω supports the controllability estimation process by tracking the expected differ ence P(s'| s,a) − P(s'| s) dynamically (or, equivalently, $\delta_{SS'} - \delta_{SAS'}$). The logic of this process is that, in a controllable environment, actions contribute to predicting the upcoming states and therefore P(s'| s,a) > P(s'| s). Higher values of Ω therefore imply higher evidence that the environment is controllable. The update of Ω is governed by the following equation:

$$\Omega \leftarrow \Omega + \alpha_{\Omega}(P(s'|s,a) - P(s'|s) - \Omega)$$

Where $\alpha\Omega \in [0,1]$ is the learning rate controlling to which extent Ω is determined by the most recent observations.

Since Ω reflects the causal influence of one's action over state transition, it can be used as a proxy to infer whether the environment is likely controllable or uncontrollable. In order to form the arbitration term reflecting this inference and accommodate inter individual differences at this step, Ω is thus transformed using a parametrized sigmoïd function:

$$\omega = \frac{1}{1 + \exp(-\beta_{\Omega}(\Omega - thres_{\Omega})}$$

Where $thres_{\Omega} \in [-1,1]$ corresponds to the threshold above which Ω is interpreted as evi dence that the environment is controllable and where $\beta_{\Omega} \in [0,Inf\ ]$ determines to which extent evidence that the environment is controllable (i.e. Ω − thresΩ > 0) favors reliance on learned SAS' transitions when making predictions (and vice-versa for SS' transitions when Ω−$thres_{\Omega}$ < 0). Thus, the variable ω implements the arbitration between the "actor" and the "spectator" model.

When only SAS' learning is considered, the probability that a given state S'=i will be observed given S and A is directly given by:

$$p(S' = i) = p(S' = i|S, A)$$

When the SS'-SAS'-Ω architecture is used, the probability that a given state S'=i will be observed given S, A and ω is directly given by:

$$p(S' = i) = \omega \max_{j=1:3} p(S' = i|S_j, A) + (1 - \omega)p(S' = i|S)$$

The max operation reflects the fact that participants are explicitly instructed that, in this version of the explore-and-predict task, their actions have the same consequences independently of the state in which they are. Thus, it is reasonable to expect that, under the hypothesis that the environment is controllable, participants will select the most likely transition independently of the state in which they are. Obviously, this step cannot be implemented if only SAS' learning is used, as the model would then lose the ability to discriminate amongst different states.

The probability that the participant predicts the next state would be i (e.g. a square) when confronted to the hypothetical state-action pair S,A (e.g. circle state, blue action) is given by:

$$p(prediction = i) = \frac{\exp(\beta_{choice} p(S' = i)}{\sum_{j=1}^{j=3} \exp\left(\beta_{choice} p(S' = i)\right)}$$

Where $\beta_{choice} \in [-Inf, Inf]$ determines to which extent the participants will systematically select the most likely transition (i.e. the highest p(S'=i), according to what has been learned) to make their predictions. A very positive β_{choice} implies that the participant systematically selects this most likely transition. A β_{choice} around 0 implies that the participant mostly makes random guesses. And a β_{choice} very negative would imply that the participant mostly goes against what he/she has learned.

The full model space was composed of SAS' alone, SAS'-SS'-Ω balanced and symmetric, SAS'-SS'-Ω unbalanced and symmetric, SAS'-SS'-Ω balanced and asymmetric. SAS'-SS'-Ω unbalanced and asymmetric. Unbalanced algorithms refer to variants allowing different learning rates for the actor and spectator models. Asymmetric algorithms refer to variants allowing different learning rates for upward and downward updates of Ω. The last 3 algorithms were used to test for imbalances or asymmetries which could contribute to biased perceptions of controllability (e.g. illusion of control), independent from Ω. A full

description of these models and of a second model space using a different learning logic is available in Supplemental Information.

Model fitting procedures

Model fitting was performed using a Variational Bayesian (VB) estimation procedure using the well-validated VBA toolbox(Daunizeau et al., 2014). The fitting procedure only attempted to explain decisions made in prediction trials. In other words, the decisions made in exploratory trials only indirectly constrained the fit by determining the information gleaned between pairs of prediction trials. For the behavioural experiments, the prior distributions of the various learning rates and of the threshold parameter were innately defined as Gaussian distributions of mean 0 and variance 3, which approximates the uniform distribution over the interval of interest after sigmoid transformations. The prior distributions of βchoice and βω parameters were defined as Gaussian distribution of mean 0 and variance 10. For the fMRI and the stress experiments, the prior distributions of every parameter was defined using the posterior mean and variance obtained from the 50 participants who passed the behavioural experiment. Hidden states (i.e values) corresponding to transition probabilities were systematically initialized at 1/3 (equiprobability prior), while Ω was initialized at 0. Detailed information about parameter transformation, model fitting, model comparison and simulation procedures is available in Supplemental Information.

fMRI: acquisition

All images were collected using a 3T Siemens Magnetom Prismafit MRI scanner (Erlangen, Germany) with a 32-channel head coil. A T2*-weighted multiband echo planar imaging sequence with acceleration factor 8 (MB8) was used to acquire BOLD-fMRI whole-brain covered images (TR = 700 ms, TE = 39 ms, flip angle = 52, voxel size = 2.4 × 2.4 × 2.4 mm3 , slice gap = 0 mm, and FOV = 210 mm). This state-of-the-art sequencing protocol was optimized from the recommended imaging guidelines of the Human Connectome Project, with the fast acquisition speed facilitating the detection and removal

of non-neuronal contributions to BOLD changes (http://protocols.humanconnectome.org/HCP/3T/imaging-protocols.html).

The experiment was divided in 4 blocks lasting on average 7.7+/-2.1 minutes (662+/-179 volumes). We recorded participants' heartbeats using the scanner's built-in photoplethysmograph, placed on the right index finger. Respiration was measured with a pneumatic belt positioned at the level of the abdomen. Anatomical images were acquired using a T1-weighted MPRAGE sequence, using a GRAPPA acceleration factor of 2 (TR = 2300ms, TE = 3.03 ms, voxel size = 1x1x1mm, 192 transversal slices, 8° flip angle). Field magnitude and phase maps were also acquired.

fMRI: preprocessing

fMRI data processing and statistical analyses were performed using statistical parametric mapping (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK). For each session, the first 4 volumes were automatically discarded by the scanner. Functional images were slice-time corrected, unwarped using the field maps and realigned to the mean functional image using a rigid-body registration. Functional images were then coregistered to the anatomical T1. Next, the anatomical images were segmented based on tissue prior probability maps for spatial normalisation using the DARTEL algorithm and the resulting normalization matrix was applied to all functional images. Finally, all images were spatially smoothed with a 6mm Gaussian kernel, except in the decoding analysis for which unsmoothed images were used.

fMRI analyses

Statistical analyses of fMRI signals were performed using a conventional two-levels random effects approach in SPM12. All general linear models (GLM) described below included the 6 unconvolved motion parameters from the realignment step. We also included the eigenvariate of signals from cerebrospinal fluid (CSF) in our GLM (fourth and lateral ventricular). Moreover, we used a retrospective image correction (RETROICOR)

method to regress out physiological noise, using 10 cardiac phase regressors and 10 respiratory phase regressors obtained by expanding cosines and sines of each signal phase to the 5th order. We also included time shifted cardiac rates (lag: +6, +10 and +12s) and respiratory volume (-1 and +5s) as nuisance regressors. All regressors of interest were convolved with the canonical hemodynamic response function (HRF). All GLM models included a high-pass filter to remove low-frequency artifacts from the data (cut-off = 96s) as well as a run-specific intercept. Temporal autocorrelation was modeled using an AR(1) process. All motor responses recorded were modeled using a zero-duration Dirac function. We used standard voxel-wise threshold to generate SPM maps ($p<0.001$ uncorrected), unless notified otherwise. All statistical inferences based on whole-brain analyses satisfied the standard multiple comparison threshold ($p(FWE)<0.05$) at the cluster level unless notified otherwise. Prediction error and (log-transformed) decision time regressors were systematically z-scored to exclude scaling effects.

All GLM models included separate onset regressors for motor responses, for prediction trials and for the first trial of each exploratory sequence (where no prediction error was elicited). All models also included parametric regressors for reaction time and ω (reflecting controllability estimates) on prediction trials. A detailed description of the GLMs used to analyze neuroimaging data is available in Supplemental Information. These GLMs only differ in the way exploratory trials were treated.

In order to verify the robustness of our whole-brain results and inspect the time course of our parametric effects of interest, we performed mixed-effects analyses on BOLD signal filtered and adjusted for nuisance regressors. This adjusted signal was extracted from the functional clusters uncovered by whole-brain analyses and segmented into trial epochs from -3 to +16 seconds around the onset of each exploration trial (excluding the first of each streak). We then estimated the effect of each regressor of interest, at each time point, for all subjects simultaneously. Subject identity was included as a random effect and a subject-specific intercept was included. Parametric regressors were z-scored in the same way as in the mass univariate anal yses. Importantly, this approach was not used

for statistical inference — since doing so would constitute double-dipping — but merely for visualization purposes.

Decoding analyses were performed using the TDT toolbox (Hebart et al., 2015). Each mini-block of 6 exploratory trials was arbitrarily coded as +1 (controllable) or -1 (uncontrollable) based on the rule governing transitions. We used a leave-one-run out cross-validation scheme with 100 permutations per subject, so that classes remained balanced for training. Training was performed on the beta values associated with each miniblock (see previous section) using a Support Vector Machine (SVM) classifier (L2-loss function, cost parameter set to 1, Liblinear, version 1.94), without feature selection or feature transformation. Since we did not constrain the testing sets to have balanced classes, balanced accuracies were used when reporting the results of the searchlight analysis (12mm sphere) at the whole-brain level.

Statistical procedures

Model selections relied on Bayesian model comparisons and exceedance probabilities, as implemented by the VBA toolbox (Daunizeau et al., 2014). The analysis of predictive accuracies relied on a 2-way repeated measure ANOVAs or one-sample t-tests, assuming normal distribution of the data following *arcsin* transformation. The analysis of decision times was performed in two steps: first, a logistic regression was performed on binarized decision times (median-split) ; second, group-level significance was assessed by means of one-sample t-tests. For the analysis of decision times, we excluded trials in which decision times were 3 standard deviations above the mean. Comparison between conditions relied on paired t-tested and comparison between groups (stress experiments) relied on two-sample t-tests, unless normality assumptions were violated, in which case non-parametric equivalents were used (Wilcoxon signed rank and rank sum tests, respectively). All t-tests were two-sided unless notified otherwise. Correlations were based on Pearson coefficients unless normality assumptions were violated, in which case Spearman rank coefficients were used. Statistical procedures related to the fMRI data are described in the fMRI analysis section above.

ACKNOWLEDGEMENTS

We are grateful to Michael Frank for his constructive comments on the manuscript and computational models. We thank Paul Gaalman for his help with fMRI data acquisition. We thank Korina Elefteriadou, Fili Dianellou, Kristin Koelbel, and Julia Breen and SOLO labsupport Leiden University for their help and support in the acquisition of the data for the stress experiment. **Funding**: This work was supported by grants from the Fyssen Foundation and the behaviour and Brain Research foundation awarded to RL (Young Investigator 2017) and a Vici award from the Netherlands Organisation for Scientific Research to RC (NWO 453-14-005).

AUTHOR CONTRIBUTIONS

Conceptualization: RL, RC, VL. Methodology: RL, VL. Software & formal analysis: RL. Investigation: RL, VL. Resources: RC, ZM. Data Curation: RL, VL. Writing - Original Draft: RL. Writing - Review & Editing: RL, ZM, VL, RC. Visualization: RL. Funding acquisition: RL, ZM, VL, RC.

DECLARTION OF INTERESTS

The authors declare no competing interests.

RESOURCE AVAILABILITY

The behavioural and fMRI data used to generate the figures will be made available upon publication at the following address: https://github.com/romainligneul/controllability. Second-level SPM images will be made available upon publication at the following address: https://identifiers.org/neurovault.collection:8810. Raw behavioral and fMRI data will be made available upon request to the corresponding authors.

The scripts used to collect and analyze data will be made available upon publication at the following address https://github.com/romainligneul/controllability.

REFERENCES

Albares, M., Lio, G., Criaud, M., Anton, J.-L., Desmurget, M., and Boulinguez, P. (2014). The dorsal medial frontal cortex mediates automatic motor inhibition in uncertain contexts: Evidence from combined fMRI and EEG studies. Hum. Brain Mapp. *35*, 5517–5531.

Alvarez, R.P., Kirlic, N., Misaki, M., Bodurka, J., Rhudy, J.L., Paulus, M.P., and Drevets, W.C. (2015). Increased anterior insula activity in anxious individuals is linked to diminished perceived control. Transl. Psychiatry *5*, e591–e591.

Amat, J., Baratta, M.V., Paul, E., Bland, S.T., Watkins, L.R., and Maier, S.F. (2005). Medial prefrontal cortex determines how stressor controllability affects behavior and dorsal raphe nucleus. Nat. Neurosci. *8*, 365–371.

Barnett, L., Barrett, A.B., and Seth, A.K. (2009). Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. Phys. Rev. Lett. *103*, 238701.

Bastos, A.M., Usrey, W.M., Adams, R.A., Mangun, G.R., Fries, P., and Friston, K.J. (2012). Canonical Microcircuits for Predictive Coding. Neuron *76*, 695–711.

Bland, S.T., Hargrave, D., Pepin, J.L., Amat, J., Watkins, L.R., and Maier, S.F. (2003). Stressor Controllability Modulates Stress-Induced Dopamine and Serotonin Efflux and Morphine-Induced Serotonin Efflux in the Medial Prefrontal Cortex. Neuropsychopharmacology *28*, 1589–1596.

Bräscher, A.-K., Becker, S., Hoeppli, M.-E., and Schweinhardt, P. (2016). Different Brain Circuitries Mediating Controllable and Uncontrollable Pain. J. Neurosci. Off. J. Soc. Neurosci. *36*, 5013–5025.

Challis, C., Boulden, J., Veerakumar, A., Espallergues, J., Vassoler, F.M., Pierce, R.C., Beck, S.G., and Berton, O. (2013). Raphe GABAergic Neurons Mediate the Acquisition of Avoidance after Social Defeat. J. Neurosci. *33*, 13978–13988.

Cheng, C., Cheung, S.F., Chio, J.H., and Chan, M.-P.S. (2013). Cultural meaning of perceived control: A meta-analysis of locus of control and psychological symptoms across 18 cultural regions. Psychol. Bull. *139*, 152–188.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLoS Comput. Biol. *10*, e1003441.

Dickerson, S.S., and Kemeny, M.E. (2004). Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. Psychol. Bull. *130*, 355–391.

Diener, C., Kuehner, C., Brusniak, W., Struve, M., and Flor, H. (2009). Effects of stressor controllability on psychophysiological, cognitive and behavioural responses in patients with major depression and dysthymia. Psychol. Med. *39*, 77–86.

Dorfman, H.M., and Gershman, S.J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. Nat. Commun. *10*, 5826.

Garrison, J., Erdeniz, B., and Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. *37*, 1297–1310.

Gilboa, A., and Marlatte, H. (2017). Neurobiology of Schemas and Schema-Mediated Memory. Trends Cogn. Sci. *21*, 618–631.

Gillan, C.M., Morein-Zamir, S., Durieux, A.M.S., Fineberg, N.A., Sahakian, B.J., and Robbins, T.W. (2014). Obsessive–compulsive disorder patients have a reduced sense of control on the illusion of control task. Front. Psychol. *5*.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron *66*, 585–595.

Grogan, J.P., Sandhu, T.R., Hu, M.T., and Manohar, S.G. (2020). Dopamine promotes instrumental motivation, but reduces reward-related vigour. ELife *9*, e58321.

Haggard, P. (2017). Sense of agency in the human brain. Nat. Rev. Neurosci. *18*, 196–207.

Hahn, A., Breakspear, M., Rischka, L., Wadsak, W., Godbersen, G.M., Pichler, V., Michenthaler, P., Vanicek, T., Hacker, M., Kasper, S., et al. (2020). Reconfiguration of functional brain networks and metabolic cost converge during task performance. ELife *9*, e52443.
Hamid, A.A., Frank, M.J., and Moore, C.I. (2019). Dopamine waves as a mechanism for spatiotemporal credit assignment (Neuroscience).
Hammack, S.E., Cooper, M.A., and Lezak, K.R. (2012). Overlapping neurobiology of learned helplessness and conditioned defeat: Implications for PTSD and mood disorders. Neuropharmacology *62*, 565–575.
Harrow, M., Hansford, B.G., and Astrachan-Fletcher, E.B. (2009). Locus of control: Relation to schizophrenia, to recovery, and to depression and psychosis — A 15-year longitudinal study. Psychiatry Res. *168*, 186–192.
Hebart, M.N., Görgen, K., and Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinformatics *8*.
Huys, Q.J.M., and Dayan, P. (2009). A Bayesian formulation of behavioral control. Cognition *113*, 314–328.
Jurado-Parras, M.T., Gruart, A., and Delgado-Garcia, J.M. (2012). Observational learning in mice can be prevented by medial prefrontal cortex stimulation and enhanced by nucleus accumbens stimulation. Learn. Mem. *19*, 99–106.
Kerr, D.L., McLaren, D.G., Mathy, R.M., and Nitschke, J.B. (2012). Controllability Modulates the Anticipatory Response in the Human Ventromedial Prefrontal Cortex. Front. Psychol. *3*.
Kim, D., Park, G.Y., O′Doherty, J.P., and Lee, S.W. (2019). Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. Nat. Commun. *10*, 5738.
Klein-Flügge, M.C., Wittmann, M.K., Shpektor, A., Jensen, D.E.A., and Rushworth, M.F.S. (2019). Multiple associative structures created by reinforcement and incidental statistical learning mechanisms. Nat. Commun. *10*, 4835.
Kühn, S., Brass, M., and Haggard, P. (2013). Feeling in control: Neural correlates of experience of agency. Cortex *49*, 1935–1942.
Kumaran, D., Warren, D.E., and Tranel, D. (2015). Damage to the Ventromedial Prefrontal Cortex Impairs Learning from Observed Outcomes. Cereb. Cortex *25*, 4504–4518.
Lee, S.W., Shimojo, S., and O'Doherty, J.P. (2014). Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. Neuron *81*, 687–699.
Leech, R., and Sharp, D.J. (2014). The role of the posterior cingulate cortex in cognition and disease. Brain *137*, 12–32.
Leibfried, F., Pascual-Diaz, S., and Grau-Moya, J. (2020). A Unified Bellman Optimality Principle Combining Reward Maximization and Empowerment. ArXiv190712392 Cs Stat.
Li, Y.S., Nassar, M.R., Kable, J.W., and Gold, J.I. (2019). Individual Neurons in the Cingulate Cortex Encode Action Monitoring, Not Selection, during Adaptive Decision-Making. J. Neurosci. *39*, 6668–6683.
Liljeholm, M., Wang, S., Zhang, J., and O'Doherty, J.P. (2013). Neural Correlates of the Divergence of Instrumental Probability Distributions. J. Neurosci. *33*, 12519–12527.
Ly, V., Wang, K.S., Bhanji, J., and Delgado, M.R. (2019). A Reward-Based Framework of Perceived Control. Front. Neurosci. *13*, 65.
Maier, S.F., and Seligman, M.E. (1976). Learned helplessness: Theory and evidence. J. Exp. Psychol. Gen. *105*, 3–46.
Maier, S.F., and Seligman, M.E.P. (2016). Learned helplessness at fifty: Insights from neuroscience. Psychol. Rev. *123*, 349–367.
Mohamed, S., and Jimenez Rezende, D. (2015). Variational Information Maximisation for Intrinsically Motivated Reinforcement Learning. Adv. Neural Inf. Process. Syst. *28*, 2125–2133.
Moscarello, J.M., and Hartley, C.A. (2017). Agency and the Calibration of Motivated Behavior.

Trends Cogn. Sci. *21*, 725–735.
Niv, Y. (2019). Learning task-state representations. Nat. Neurosci. *22*, 1544–1553.
O'Callaghan, C., Vaghi, M.M., Brummerloh, B., Cardinal, R.N., and Robbins, T.W. (2019). Impaired awareness of action-outcome contingency and causality during healthy ageing and following ventromedial prefrontal cortex lesions. Neuropsychologia *128*, 282–289.
O'Doherty, J. (2004). Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning. Science *304*, 452–454.
Otto, A.R., Raio, C.M., Chiang, A., Phelps, E.A., and Daw, N.D. (2013). Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. *110*, 20941–20946.
Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.
Shahar, N., Hauser, T.U., Moutoussis, M., Moran, R., Keramati, M., Consortium, N., and Dolan, R.J. (2019a). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. PLOS Comput. Biol. *15*, e1006803.
Shahar, N., Moran, R., Hauser, T.U., Kievit, R.A., McNamee, D., Moutoussis, M., NSPN Consortium, and Dolan, R.J. (2019b). Credit assignment to state-independent task representations and its relationship with model-based decision making. Proc. Natl. Acad. Sci. *116*, 15871–15876.
Shenhav, A., Cohen, J.D., and Botvinick, M.M. (2016). Dorsal anterior cingulate cortex and the value of control. Nat. Neurosci. *19*, 1286–1291.
Spengler, S., von Cramon, D.Y., and Brass, M. (2009). Was it me or was it you? How the sense of agency originates from ideomotor learning revealed by fMRI. NeuroImage *46*, 290–298.
Vejmelka, M., and Paluš, M. (2008). Inferring the directionality of coupling with conditional mutual information. Phys. Rev. E *77*.
Vesuna, S., Kauvar, I.V., Richman, E., Gore, F., Oskotsky, T., Sava-Segal, C., Luo, L., Malenka, R.C., Henderson, J.M., Nuyujukian, P., et al. (2020). Deep posteromedial cortical rhythm in dissociation. Nature *586*, 87–94.
Voss, M., Chambon, V., Wenke, D., Kühn, S., and Haggard, P. (2017). In and out of control: brain mechanisms linking fluency of action selection to self-agency in patients with schizophrenia. Brain *140*, 2226–2239.
Wanke, N., and Schwabe, L. (2019). Subjective Uncontrollability over Aversive Events Reduces Working Memory Performance and Related Large-Scale Network Interactions. Cereb. Cortex N. Y. N 1991.
Weiss, A., Chambon, V., Lee, J.K., Drugowitsch, J., and Wyart, V. (2019). Interacting with volatile environments stabilizes hidden-state inference and its brain signatures (Neuroscience).
Westbrook, A., van den Bosch, R., Määttä, J.I., Hofmans, L., Papadopetraki, D., Cools, R., and Frank, M.J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. Science *367*, 1362–1366.
Wise, T., Marwood, L., Perkins, A.M., Herane-Vives, A., Joules, R., Lythgoe, D.J., Luh, W.-M., Williams, S.C.R., Young, A.H., Cleare, A.J., et al. (2017). Instability of default mode network connectivity in major depression: a two-sample confirmation study. Transl. Psychiatry *7*, e1105.
Wood, K.H., Wheelock, M.D., Shumen, J.R., Bowen, K.H., Ver Hoef, L.W., and Knight, D.C. (2015). Controllability modulates the neural response to predictable but not unpredictable threat in humans. NeuroImage *119*, 371–381.

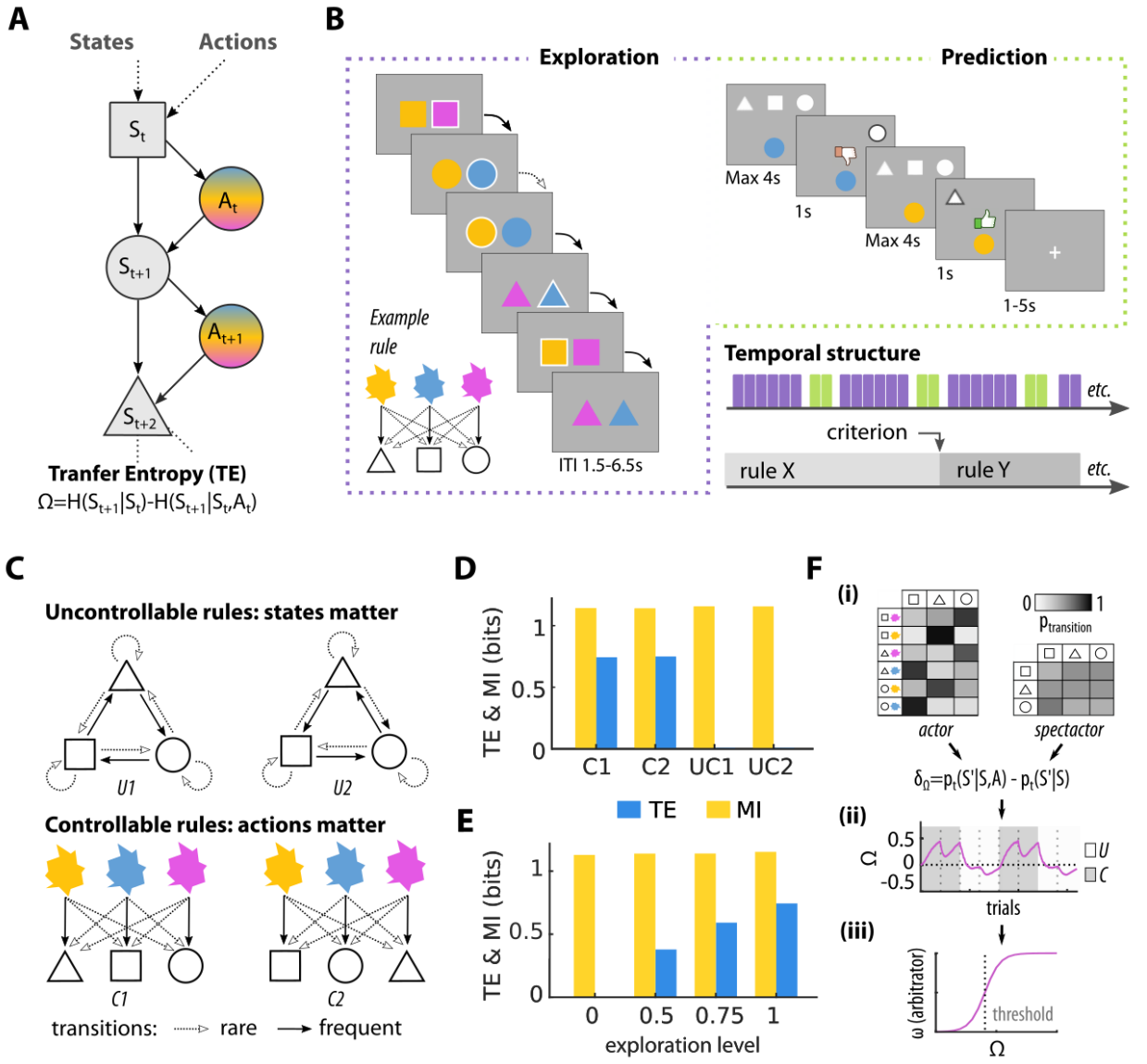

Figure 1. Theoretical framework and experimental protocol.

(A) Controllability can be inferred from transfer entropy, an information-theoretic measure quantifying to which extent a time series causally influences another one.

(B) Time course of a novel explore-and-predict task divided in short mini-blocks. Each miniblock consists of a series of exploratory trials (violet) followed by two counterfactual prediction trials (green) used to assess learning and subjective controllability.

(C) Representation of the 2 uncontrollable rules (U1, U2) and the 2 controllable (C1, C2) rules, which alternate covertly to govern the evolution of the environment.

(D) Simulations showing the dissociation of controllability, as indexed by TE, and predictability, as indexed by the mutual information (MI) shared between successive state-action pairs (random exploration policy).

(E) Simulations under a controllable rule showing that TE requires exploration to be used as a proxy for controllability. Indeed, in the absence of exploration, the conditional

entropies H(S'|S) and H(S'|S,A) are identical (see also Supplemental Text and Figure S1C).
(F) Synthetic overview of the algorithm able to derive an online approximation of TE (termed Ω) by comparing on each trial the transition probabilities of an actor (SAS') and a spectator (SS') model of the world. By thresholding Ω, the algorithm could in turn arbitrate between spectator and actor models when making predictions depending on current controllability estimates (ω).

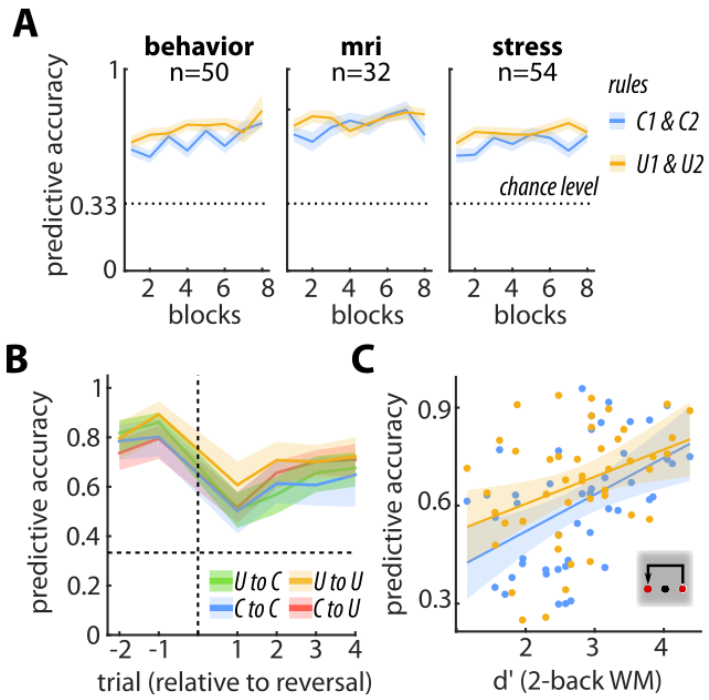

Figure 2. Behavioural performance.

(A) Accuracy in the prediction trials for each of the 3 experiments, split by condition (see also Table S1).

(B) Fast recovery of predictive accuracy around reversals.

(C) Positive correlation between working memory capacity indexed by a 2-back task (see Supplemental Information) and accuracy in the explore-and-predict task for controllable (blue) and uncontrollable (orange) contexts.

Shaded areas represent standard errors of the mean (SEM).

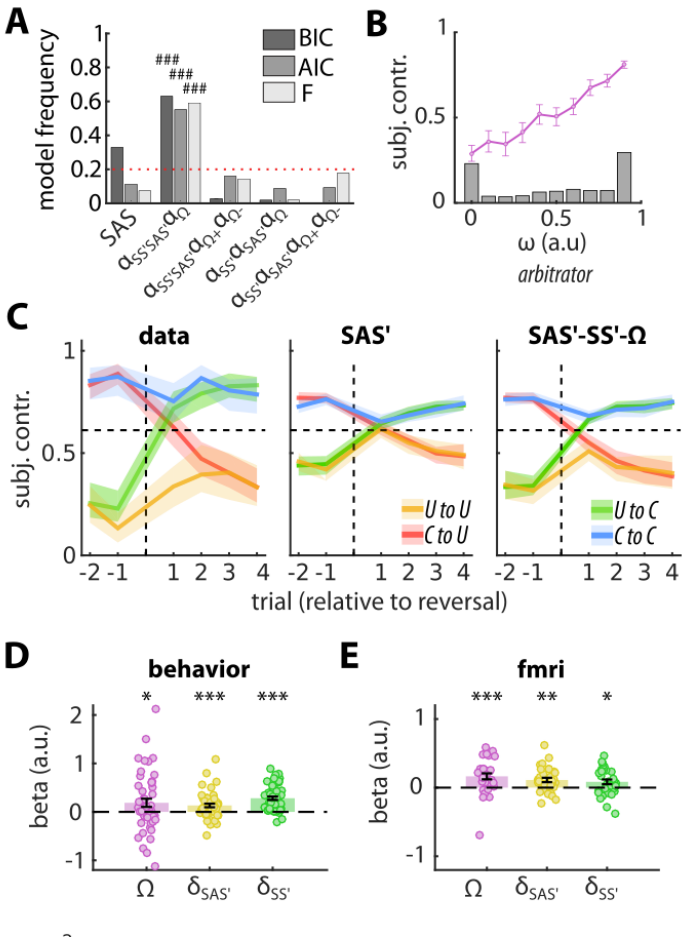

Figure 3. Computational modeling.

(A) Bayesian group model comparison pooled over three experiments showed the advantage of the simplest controllability scheme (SS'-SAS'-Ω) over the actor model (SAS') alone (see Methods and Figure S2A and 3).

(B) Normalized distribution of the arbitrator variable ω (grey bar) and its linear relationship with subjective controllability (pink line). Pairs of prediction trials were labelled as "subjectively controllable" when counterfactual predictions differed (e.g. different responses for blue and yellow actions in the circle state).

(C) The SS'-SAS'-Ω scheme better captured the dynamics of subjective controllability around reversals.

(D-E) Coefficients of the logistic regression predicting binarized reaction times in the behavioural and fMRI experiment using actor and spectator prediction errors (δSAS' and δSS') and Ω.

Error bars and shaded areas represent SEM. *p<0.05,***p<0.005, ***p<0.001. ###$p_{exceedance}$ > 0.999. BIC: Bayesian Information Criterion. AIC: Akaike Information Criterion. F: Free Energy.

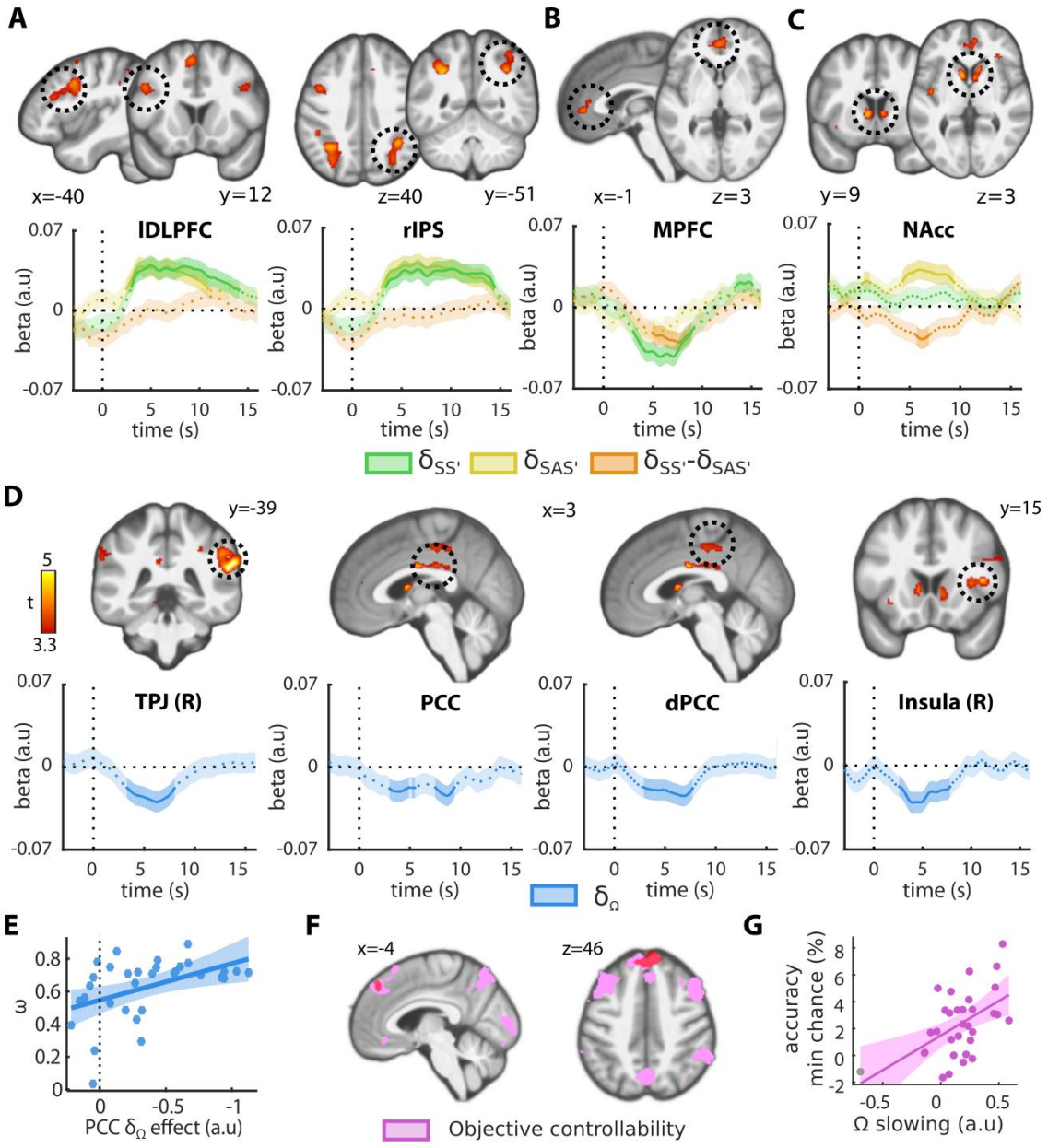

Figure 4. Neural dissociation of the actor and spectator models during exploration. (A) Brain regions whose activity was higher when the prediction error terms ($\delta_{SS'}$ and $\delta_{SAS'}$) were both above their median value, as compared to when they were both below. (B) Paired t-test showing the brain areas whose activity dissociated in response to trials where only $\delta_{SAS'}$ was above median, compared with those where only $\delta_{SS'}$ was above median. This analysis revealed that the mPFC encoded specifically $\delta_{SS'}$ but not $\delta_{SAS'}$. (C) Parametric analysis of BOLD responses showed that the mPFC and the nucleus accumbens encoded negatively the difference term $\delta_{SS'}$ - $\delta_{SAS'}$ used to update controllability. Contrasting with the mPFC pattern, $\delta_{SAS'}$ was encoded positively In the nucleus accumbens whereas $\delta_{SS'}$ was not.

(D) Brain regions encoding signed the second-order prediction errors δ_Ω. All areas surviving correction for multiple comparison showed a negative effect, implying greater activity when an action was less causal than expected.
(E) The degree to which PCC encoded δ_Ω predicted the propensity to rely on the actor model across participants.
(F) Decoding of controllability (rule type) from brain data. A searchlight analysis revealed that the dmPFC, the dlPFC, the right TPJ and the precuneus were sensitive to environmental controllability.
(G) The sensitivity of the dmPFC to objective controllability predicted to which extent periods of higher controllability led to slower decision times.
The time courses shown below (A-D) were only used for robustness checks and visualization. Statistical inferences were based on whole-brain effects at standard thresholds (voxel-wise: $p<0.001$, uncorrected; cluster-wise: $p<0.05^{FWE}$). Shaded areas represent SEM.

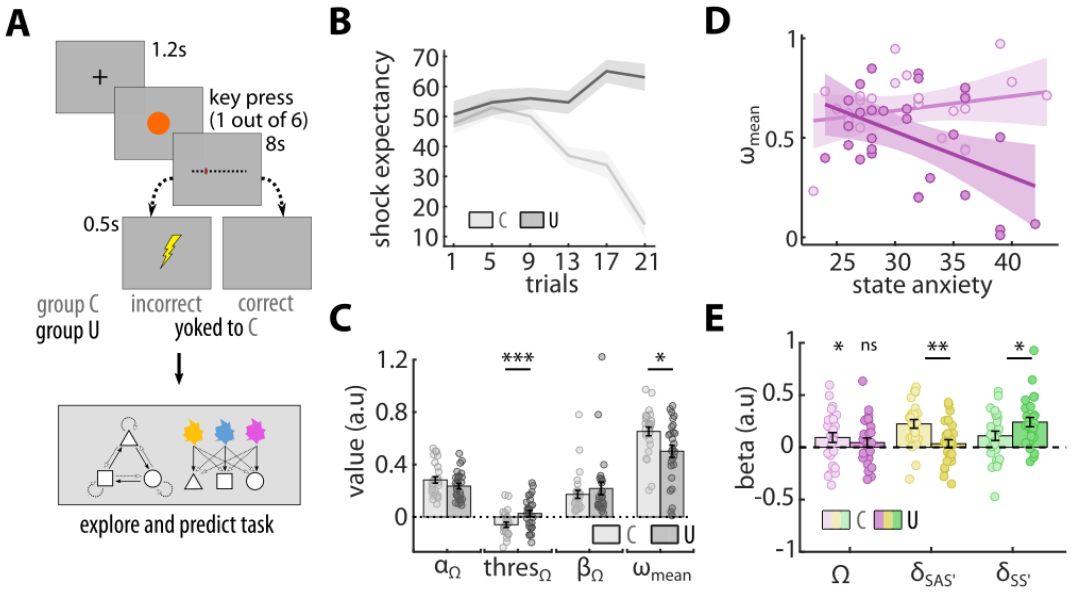

Figure 5. Stress experiment.

(A) Induction of controllable and yoked uncontrollable stress followed by the explore-and-predict task.

(B) Temporal evolution shock expectancy during the induction phase, split by condition.

(C) Impact of induction type on best-fitting parameters related to controllability monitoring as well as on the mean value taken by the arbitrator ω in the explore-and-predict task. The threshold parameter $thres_\Omega$ which determined above which value Ω was treated as evidence for a controllable environment increased significantly following uncontrollable stressors and the average value of the arbitrator was reduced, hence leading to increased reliance on the spectator model when making predictions (β_Ω and α_Ω were unaffected, β_Ω was divided by 100 for display purposes).

(D) State anxiety moderated the effect of induction type on the arbitrator variable (ω) reflecting controllability estimation. Higher state anxiety was associated with greater reliance on the spectator model after exposure to uncontrollable stressors.

(E) The impact of induction type on the slowing of decision times induced by $\delta_{SS'}$, $\delta_{SAS'}$ and Ω was consistent with increased reliance on the spectator model. Indeed, the effect of Ω and $\delta_{SAS'}$ vanished whereas the effect of $\delta_{SS'}$ increased in the uncontrollable conditions.

All error bars and shaded areas represent SEM. *p<0.05,**p<0.01.

Methods and Supplemental Information for

Stress-sensitive brain computations of task controllability

Romain Ligneul, Zachary Mainen, Verena Ly, Roshan Cools

Correspondence to:
romain.ligneul@research.fchampalimaud.org
roshan.cools@fcdonders.ru.nl

This PDF file includes:

Methods

1 Participants

1.1 Behavioral experiment

Fifty young adult participants (mean age: 24.7, range: 18—43, 27 women) were recruited via the Sona system (human subject pool management system) of the Radboud University (The Netherlands). Participants were prescreened to ensure normal to corrected vision, a sufficient understanding of English language, no history of neurological of psychiatric diseases. Pregnancy and use of psychoactive drugs in the 2 weeks prior to the experiments were part of the exclusion criteria. All participants were included in the data analysis. They were compensated at a fixed rate of 8 euros/hour for their participation in the study. The study was approved by the local ethics committee (CMO region Arnhem/Nijmegen, The Netherlands, CMO2001/095) and all participants provided written informed consent, in line with the declaration of Helsinki.

1.2 fMRI experiment

Thirty-two young adult participants (mean age: 25.1, range: 20—43, 18 women) were recruited via the human subject pool management system of the Radboud University (The Netherlands). Participants were pre-screened and excluded using the same criteria than for the behavioral experiment. One participant was excluded *a posteriori* from the study due to excessive sleepiness in the scanner (eyes-closed more than 10 % of the time n=1). All participants were compensated at a fixed rate of 10 euros/hour for their participation in the study. Most participants were also included in the behavioral experiment or in an previous pilot study (n=29) but inclusion in the fMRI experiment was independent of performance in the behavioral experiment. The study was approved by the local ethics committee (CMO region Arnhem/Nijmegen, The Netherlands, CMO2001/095) and all participants provided written informed consent, in line with the declaration of Helsinki.

1.3 Stress experiment

A total of 62 participants (mean age = 21.8; range: 18-27, 52 women) were recruited via SONA, the human participants system, at Leiden University. Exclusion criteria were: pregnancy, deviation from normal or corrected-to-normal vision, insufficient understanding of written and spoken English, self-reported history of any neurological, cardiac, or psychiatric disease, any use of prescribed psychotropic drugs within two weeks or alcohol in the 24 hours prior to the experiment. The study was approved by the Psychology Research Ethics Committee (CEP17-0905/282) at Leiden University. All participants provided written informed consent and were compensated for their participation: 7.50 euros/hour or 2 participant credits. Four participants needed to be excluded for different reasons: 1. no aversive stimuli being delivered due to cable being connected to the wrong port; 2. stressor manipulation task crashed during the experiment; 3. unable to finish the experiment due to delays; 4. noncompliance to instructions of the experimenter. To retain the yoked design, we also excluded the four yoked counterparts of these participants. For another four participants in the uncontrollable condition, an error occurred in the yoking procedure resulting in one aversive stimulus less for these participants. Nevertheless, we decided to include these participants and their yoked counterparts in the analyses: if anything, this single extra aversive stimulus in the controllable condition would only make the comparison with the uncontrollable condition more conservative. Thus, after exclusion of 8 participants, a total sample size of 54 was left for analyses.

2 Explore-and-predict task

2.1 Instructions and training

Upon arrival on site and after completing the informed consent form, participants were told that they would take part to an experiment probing the human ability to detect, learn and use

the rules which govern a simplified environment composed of three states and three actions. The experimenter explained that states corresponded to the geometrical shapes displayed on the screen (triangle, circle, square), whereas the actions related to the colors of these shapes (blue, yellow, magenta). Participants were told that they would have to explore this simplified environment by selecting, in each state, an action using the left/right arrows of the keyboard (behavioral and stress experiments) or button box (fMRI experiment). They were then presented with a scheme, similar to that of Fig. 1C, describing the 4 possible rules which controlled the transitions from one state to another.

The experimenter explained the distinction between controllable and uncontrollable rules and told participants that their goal would be to learn the rule during exploration and that, from time to time, the computer program would ask them to predict the next most likely state given a hypothetical state-action pair. The experimenter explained the logic of counterfactual testing, making sure that participants would understand the implications of making the same prediction for the two possible actions associated with a given state (i.e no decisive influence of actions implying a "spectator" rule); versus making distinct predictions (i.e. decisive influence of actions implying an "actor" rule).

The participants were told that rules would alternate covertly, without warning, and that they should thus constantly evaluate whether the rule they learned was still valid. They were also informed that the number of trials in the task would depend on their ability to make explicit predictions, but the exact criterion determining covert reversals was not revealed.

Following these verbal instructions, participants completed a training phase divided in 4 blocks (behavioral) or 8 blocks (stress and fMRI experiments). At the beginning of each block, the rule which governed the transitions was displayed explicitly on the screen. Participants were told to "feel what it is like to explore the environment when such a rule is active" and they were given the opportunity to make predictions from time to time, as in the testing phase. Positive

or negative feedback was delivered following each prediction. Once the participants had completed all training blocks, the experimenter verified that they had understood the principle of the experiment. Next, they explained that, unlike the training phase, the transitions of the test phase would be slightly noisy. This noise was described as a small proportion of trials in which the computer program would select randomly the next state instead of applying the active rule. It was made explicit that, in the prediction trials, the noise would be absent and that the feedback - if any - would faithfully reflect the active rule.

The relatively long duration of the training phase (15-20 minutes) was needed to ensure that most participants would get the principle of the task. A pilot study had suggested that only a subset of participants were able to operate successfully in the task with minimal instructions and training. Note that participants in the fMRI experiment completed the training phase twice, once outside the scanner (4 blocks) and once inside the scanner (4 blocks) during the acquisition of the anatomical scan.

2.2 Structure of the task

In the 3 experiments, the overall structure of the task was identical. Participants performed 6 (fMRI and stress experiment) or 7 (behavioral experiment) exploratory trials before a pair of predictions was required. Pairs of predictions always probed the two actions available for a given state (e.g blue followed by yellow in the circle state), in order to derive subjective controllability from counterfactual responses. Participants received feedback about their predictions in 50 percent (fMRI and stress) or 100 percent (behavioral experiment) of the trails. In the fMRI and stress tasks, feedback was delivered only after one of the two counterfactual predictions in order to prevent participants from inferring whether the rule was controllable or not based on prediction feedback.

On each exploratory trial, two identical geometrical shapes were displayed side by side. The

visual angle encompassing both shapes was about 6-7° for all experiments. The color of each shape determined the action corresponding to left and right button presses (the side was randomly assigned in each trial). In the behavioral experiment, these shapes were displayed for at least 0.5s or until the participant made a choice. The shapes started to fade out automatically after 1.1s in the fMRI experiment and the fading was complete after 1.5s (in the behavioral experiment, shapes faded out at button press in 0.25s). In every case, a small warning symbol urging participants to make a choice appeared between the two shapes if no button press had been registered after 1.5s. followed by a blank screen whose duration was:

- Uniformly distributed in the [0.25, 0.5] interval (mean 0.375s) for the behavioral experiment
- Exponentially distributed in the [0.05 5] interval (mean 2s) for the fMRI experiment.
- Exponentially distributed in the [0.05 4.5] interval (mean 0.5s) for the stress experiment.

The first prediction trial of each pair was simply displayed at the end of the ITI of the previous exploratory trial. These trials were self-paced, although a warning appeared on screen after 4s to urge participants to make a choice. The hypothetical state action pair was displayed at the center of the screen (visual angle of about 2°), just below a question mark, and the 3 possible next states were displayed as white geometrical shape at the top of the screen. The selected state was then highlighted for 0.3s. In the next period of 0.7s the feedback was displayed (if no feedback was displayed, the selected state simply remained highlighted during 0.7s). In the behavioral experiment, a blank screen lasting 0.3s followed each prediction trial. In the fMRI and stress experiements, it was followed by a blank screen whose duration was distributed similarly to the ITI describe above (0.05 to 4.5 or 5, respectively).

2.3 Reversal schedule

The ongoing rule was never changed before 4 pairs of predictions were completed. In the behavioral experiment, the rule changed from then as soon as 5 correct responses were provided in the last 6 predictions or if the last 4 predictions were accurate. In the fMRI experiment, the rule was changed as soon as the p-value of a binomial test indicated that accuracy was significantly below chance ($p<0.05$, one-tailed, chance level: 1/3), hence making the accuracy threshold more lenient as the number of predictions made for a given rule increased. In all experiments, the rule changed after 10 pairs of predictions, even if performance did not meet the learning criterion. Inability to reach this criterion could be due to distorted controllability perception or simply a slow, suboptimal learning process. Therefore, We did not exclude any data based on performance. Predictions were pseudo-randomly ordered with the constraint that each state would be tested a similar number of times.

Contrary to the stress and fMRI experiments which were divided in 4 identifiable blocks separated by short pause screen showing a slide describing the possible rules, the version of the SS'SAS' task used for the behavioral experiment was not explicitly divided into blocks. In all experiments, the computer program covertly alternated rules to enable studying changes in predictions following 4 uncontrollable (U) to controllable (C) reversals, 4 C to U reversals, 2 C-C reversals (e.g. from rule C1 to rule C2) and 2 U-U reversals (e.g. U1 to U2). Because inter-block transitions were not taken into considerations (the participants were informed that each block was independent), stress and fMRI experiments thus required 4 blocks in which the 4 rules were tested (3 transitions per blocks). Block types were counterbalanced across participants. The 4 possible blocks were: $\{U1, U2, C2, C1\}$, $\{U2, C1, U1, C2\}$, $\{C1, C2, U2, U1\}$ and $\{C2, U1, C1, U2\}$, hence resulting in 8 cross-dimensional (i.e U to C, C to U) and 4 intra-dimensional (i.e U to U, C to C) reversals. This ordering enabled us to distinguish amongst different computational models and to promote the use of a controllability monitoring strategy.

Note that each rule was tested once in each position within blocks.

2.4 Transition noise

Finally, in order to increase the variance and the temporal distribution of prediction errors related to $p(S'|S)$, $p(S'|S,A)$ and Ω, the transitions were noisy. In the behavioral experiment, a random noise level was set to 7.5% (so that the active rule would be applied in 92.5% of the exploratory trials). In the fMRI experiment, the noise level depended on the transition. Within each rule, the three possible SA-S' or S-S' transitions had thus a noise of 5, 10 or 20% (e.g. under U1, square to circle would be realized 95% of the time while circle to triangle would be realized only 80% of the time). This refined procedure was used to increase the variance of the prediction error terms regressed onto neural activity without completely disrupting learning in less efficient participants (the lowest noise transitions being still easily detectable). Finally, in the stress experiment, the noise level was set to 10% for all transitions.

3 Other tasks

3.1 Working memory task

The working memory (WM) task was programmed using Psychtoolbox (Matlab 2014b). It tested four different conditions: 2-back, 3-back, 2-forth and 3-forth. In every condition, participants were presented with a constant stream of numbers (display duration: 700ms, inter-trial interval: 800ms. In the n-back blocks, they were instructed to press the space bar each time a number would be identical to the number encountered 2 or 3 trials before, depending on the difficulty level. In the n-forth task, participants were asked to detect streak of 3 or 4 numbers increasing or decreasing in a row (data not analyzed). Each block was composed of 50 trials and included eight targets. Incorrect responses (i.e. miss or false alarm) triggered a warning message (1s). Before the task itself, participants performed 16 trials (2 targets) of each condi-

tion. Working memory performance was assessed using a d-prime criterion (z(HIT)-z(FA)). Since the d-prime could not be computed in one participant who had difficulty completing the 3-back task, we restricted our analysis to the 2-back task. Note that 4 more participants in the behavioral experiment did not complete the working memory task due to technical problems, hence resulting in 46 participants included in WM-related analyses.

3.2 Stress induction task

To test the impact of prior controllability over stress on subsequent controllability estimations, participants underwent a stressor controllability manipulation prior to the explore-and-prediction task in a between-subjects design. Critically, we employed a between-subjects yoked control procedure in order to match the amount and order of aversive outcome stimuli between the controllable and uncontrollable conditions. We randomized participants in blocks of four (two Controllable and two Uncontrollable conditions) where the Controllable condition of a yoked pair was always administered first in order to create the schedule for the yoked counterpart in the Uncontrollable condition.

3.2.1 General procedure

Upon arrival to the laboratory, participants were briefly reminded about the experimental procedure. The study was framed such that it was not clear that the study involved a between subject manipulation of stressor controllability but rather that it concerned decision making processes and physiological measurements. After providing informed consent, electrodes for the physiological measurements were positioned on the participant skin conductance, blood pressure, electrocardiogram (ECG) and impedance cardiography (ICG). Subsequently, they filled out several self-report questionnaires (State-Trait Anxiety Inventory, Leiden Index of Depression Sensitivity-Revised, Cognitive Emotional Regulation Questionnaire, Behavioral Activa-

tion/Inhibition Scale, Barratts Impulsiveness Scale, Childhood Trauma Questionnaire), and performed a working memory task (2-Back task, see below).

After a 5 minutes baseline measure of skin conductance, blood pressure, ECG, and ICG while watching a neutral video clip, participants received instructions and training for the explore-and-predict task. Subsequently, they went through a procedure to determine the shock intensity to be used during the stressor controllability manipulation task. By providing the instructions and training for the explore-and-predict task before exposure to any shock stimuli, we prevented any effects of the stressor on the instructions and training of the explore-and-predict task. Online measures of skin conductance, blood pressure, ECG and ICG were taken during the stressor controllability task (data not analyzed). Participants were asked to indicate their levels of positive and negative affect and cognition on a visual analogue scale on different moments during the experiment (baseline, before and after the manipulation, and at the end of the experiment)

3.2.2 Stressor controllability manipulation

Electric stimuli served as stressors in the manipulation task and were delivered by a Digitimer DS7 stimulator. First, individual levels of intensity of the electric stimulus for the manipulation task were determined using a stepwise procedure in which the intensity of the stimulus was gradually increased in intervals of 0.10 mA until participants reported a 'just bearable, but not yet painful' experience of shock on a scale from 0 = not uncomfortable at all to 100 = just bearable, but not yet painful.

Depending on the condition to which participants were randomly assigned (controllable or uncontrollable condition), perceived control over the stressor was manipulated via the presence or absence of objective control (i.e. choice and action-outcome-contingency) respectively. A yoked control-design with preprogrammed pseudorandomized schedule enabled us to match the amount and order of electric stimuli between the conditions as well as to minimize interindivid-

ual variation in the manipulation.

In the controllable condition, a total of four cues (different in shape and color) were presented for at least six repetitions each following a preprogrammed pseudorandomized schedule. Participants could – supposedly – learn by trial-and-error the correct response corresponding to the cue (a key between 1 and 6) to avoid the electric stimulus. They were instructed that each cue had a unique corresponding key as correct response and that the correct response would always prevent the electric stimulus. Unbeknownst to the participants, the trial at which they could prevent the electric stimulus for a particular cue for the first time depended on a combination of the preprogrammed schedule and optimal exploration of the participant. For example, participants could prevent the electric stimulus for cue A at the third repetition (A3) for the first time, if they explored a third key on this trial that was different from the first two attempts (for A1 and A2). This third key was then assigned as the correct key corresponding to cue A for the rest of the task. If participants repeated unsuccessful attempts for a specific cue or chose correct keys assigned to other cues, they would receive an electric stimulus. Critical trials on which participants would be able prevent the electric stimulus for the first time according to the schedule were repeated until the participants arrived at a correct response. As such, all participant underwent the whole schedule with a minimum of 24 trials, and were able to acquire the correct response for each cue.

The uncontrollable condition was yoked to the controllable condition, such that participants experienced a comparable pattern of events across conditions. However, in the uncontrollable condition, participants were not able to acquire these action-outcome contingencies to prevent the shocks, but were instead instructed to press a random button (key 1 to 9) on each trial.

4 Computational modeling

The main purpose of all SAS'-SS'-Ω variants is to provide a way to dynamically estimate the causal influence of actions over state transitions by updating a variable termed Ω. In all models, S represents the previous state of the environment, A represents the previous action and S' represents the current state of the environment. The local causality estimate Ω can only be used as a proxy for controllability, which is not a property of actions but of the environment. It is this "inferred controllability" variable, termed ω, which can then be used to decide (arbitrate) whether one should make predictions using learned S-S' transitions or learned SA-S' transitions. Ω is homologous to transfer entropy (TE, which is itself a generalization of Granger causality to discrete and non-linear domains), with 3 important differences:

- Ω is not computed based on post-hoc transition probabilities (i.e after all transitions have been observed) but on transition probabilities estimated trial-by-trial (using a delta rule). This means that Ω is dynamic and not static. Unlike TE, Ω can therefore accumulate causal evidence locally in order to discriminate between periods of high controllability and periods of low controllability.

- For the same reason, Ω can take negative values. This can happen just after rule reversals or when the active rule is suddenly violated due to noise. Indeed, in such cases, S-S' transitions can momentarily appear more likely than SA-S' transitions.

- It is not log-transformed and therefore does not represent "bits" of information. Instead, it represents the expected difference between the probability of observations given previous states and actions p(S'|S,A) and the probability of observations given previous states only $(S'|S)$. Importantly, by forcing $p(S'|S,A) \geq (S'|S)$ and by tracking the expected value of $log(p(S'|S,A)) - log(p(S'|S))$ instead of $p(S'|S,A) - p(S'|S)$, Ω and TE converge towards the same values in the long-run.

In order to demonstrate that participants used a dynamic estimate of transfer entropy to solve the task, we systematically compared variants of the SAS'-SS'-Ω architecture to a standard model-based architecture tracking SA-S' transitions [1, 2]. This approach is powerful because the asymptotic performance of this latter model is identical to that of SAS'-SS'-Ω models, whose relative advantage lies in the ability to arbitrate between controllable (i.e SAS') and uncontrollable (i.e S-S') transition matrices to perform predictions. Indeed, the actor (ie. SAS') model can perfectly learn the state-state (SS') transition probabilities tracked by the spectator model. It only needs more data points, especially when the number of possible state-action pairs is high. This remark is important because it implies that comparing the SAS'-SS'-Ω architecture to the SAS' model alone constitutes a fair and stringent test.

In the following, the computational steps relevant for the entire model space are described. While the models took into account all trials providing information about transition probabilities to update hidden states, the fitting procedure only attempted to explain decisions made in prediction trials. In other words, the decisions made in exploratory trials did not constrain the values of the best-fitting parameters.

4.1 Update of SAS' transitions ("actor model")

This module tracks SA-S' transitions. Because actions are explicitly represented, it is called the "actor" module. In each exploratory trial and prediction trial followed by a feedback, this standard model-based learning module updates the transition probabilities linking one state-action pair to the newly encountered state in the following fashion:

Realized transitions:

$$P(s'|s,a) \leftarrow P(s'|s,a) + \alpha_{sas'}(1 - P(s'|s,a))$$

Unrealized transitions:

$$P(s'|s,a) \leftarrow P(s'|s,a)(1 - \alpha_{sas'})$$

Where $\alpha_{sas'} \in [0,1]$ controls to which extent learned transition probabilities are determined by the most recent transitions. Note that $1 - P(s'|s,a)$ is noted $\delta_{SAS'}$ in the main text.

4.2 Update of SS' transitions ("spectator model")

This module tracks S-S' transitions. Because actions are not represented, it is called the "spectator" module. In each exploratory trial and prediction trial followed by a feedback, this spectator module updates the transition probabilities linking one state to the newly encountered state in the following fashion:

Realized transitions:

$$P(s'|s) \leftarrow P(s'|s) + \alpha_{ss'}(1 - P(s'|s))$$

Unrealized transitions:

$$P(s'|s) \leftarrow P(s'|s)(1 - \alpha_{ss'})$$

Where $\alpha_{ss'} \in [0,1]$ controls to which extent learned transition probabilities are determined by the most recent transitions. Note that $1 - P(s'|s,a)$ is noted $\delta_{SS'}$ in the main text.

4.3 Update of Ω

This second-order module tracks the expected difference $P(s'|s,a) - P(s'|s)$ dynamically (or, equivalently, $\delta_{SS'} - \delta_{SAS'}$). The logic of this process is that, in a controllable environment, actions contribute to predicting the upcoming states and therefore $P(s'|s,a) > P(s'|s)$. At any given moment, a positive Ω therefore constitutes evidence that the environment is controllable:

$$\Omega \leftarrow \Omega + \alpha_{\Omega}(P(s'|s,a) - P(s'|s) - \Omega)$$

Where $\alpha_{\Omega} \in [0,1]$ is the learning rate controlling to which extent Ω is determined by the most recent observations.

4.4 Inference of controllability (ω) from Ω

As written above, Ω reflects the causal influence of one's action over state transition. It can therefore be used as a proxy to infer whether the environment is likely controllable or uncontrollable. In order to form the arbitration term reflecting this inference and accommodate inter individual differences at this step, Ω is thus transformed using a parametrized sigmoïd function:

$$\omega = \frac{1}{1+\exp(-\beta_\Omega(\Omega - threshold_\Omega)}$$

Where $threshold_\Omega) \in [-1,1]$ corresponds to the threshold above which Ω is interpreted as evidence that the environment is controllable and where $\beta_\Omega \in [0, Inf]$ determines to which extent evidence that the environment is controllable (i.e. $\Omega - threshold_\Omega > 0$) favors reliance on learned SAS' transitions when making predictions (and vice-versa for SS' transitions when $\Omega - threshold_\Omega < 0$).

In other words, the variable ω implements the arbitration between the "actor" and the "spectator" model.

4.5 Prediction

When only SAS' learning is considered, the probability that a given state S'=i will be observed given S and A is directly given by:

$$p(S' = i) = p(S' = i|S,A)$$

When the SS'-SAS'-Ω architecture is used, the probability that a given state S'=i will be observed given S, A and ω is directly given by:

$$p(S' = i) = \omega \max_{j=1:3} p(S' = i|S_j,A) + (1-\omega)p(S' = i|S)$$

The max operation reflects the fact that participants are explicitly instructed that, in this version of the SS'SAS' task, their actions have the same consequences independently of the state in

which they are. Thus, it is reasonable to expect that, under the hypothesis that the environment is controllable, participants will select the most likely transition independently of the state in which they are. Obviously, this step cannot be implemented if only SAS' learning is used, as the model would then lose the ability to discriminate amongst different states.

The probability that the participant predicts the next state would be i (e.g. a square state) when confronted to the hypothetical state-action pair S,A (e.g. circle state, blue action) is finally given by:

$$p(prediction = i) = \frac{\exp(\beta_{choice} p(S' = i)}{\sum_{j=1}^{j=3} \exp(\beta_{choice} p(S' = i))}$$

Where $\beta_{choice} \in [-Inf, Inf]$ determines to which extent the participants will systematically select the most likely transition (i.e. the highest p(S'=i), according to what has been learned) to make their predictions. A very positive β_{choice} implies that the participant systematically select this most likely transition. A β_{choice} around 0 implies that the participant mostly makes random guesses. And a β_{choice} very negative would imply that the participant mostly go against what he/she has learned.

As written above, the key architecture of reference to demonstrate by means of model comparison that participants tracked two sets of transition probabilities and estimated controllability to solve the task is the SAS' model alone. Indeed, such architecture has the same asymptotic performance as the SAS'-SS'-Ω architecture (which thus constitutes a generalization of the SAS' model) in stable environments.

4.6 Model space

Hereafter we describe the 5 different models subjected to model comparison procedure.

1. *SAS' model.* This model only used SA-S' transitions to predict upcoming states. Thus, it had only **2** parameters: a learning rate $\alpha_{sas'}$ and a inverse temperature β_{choice}. The

arbitrator variable ω was set to a value of 1, so that the update of p(S'|S) and Ω had no impact whatsoever.

2. *SAS'-SS'-Ω, balanced and symmetric*. This model made use of the full architecture. The update of SA-S' and S-S' transition probabilities was balanced, as the same learning rate was used for the actor and spectator modules (i.e. $\alpha_{sas'} = \alpha_{ss'}$). The update of Ω was also symmetric, as the learning rate α_{Ω} was the same when Ω increased or decreased. Counting the two parameters controlling the inference of controllability based on Ω (i.e. $thres_{\Omega}$ and β_{Ω}) and the inverse temperature parameter β_{choice}, this model had thus **5** parameters.

3. *SAS'-SS'-Ω, unbalanced and symmetric*. This model was similar in every respect to model 2, except that $\alpha_{sas'}$ and $\alpha_{ss'}$ were independent. This model had thus **6** parameters.

4. *SAS'-SS'-Ω, balanced and asymmetric*. This model was also similar in every respect to model 2, except that learning rate controlling the update of Ω was split in two independent learning rate $\alpha_{\Omega+}$ and $\alpha_{\Omega-}$, depending on whether Ω was updated upwards or downwards. This model had thus **6** parameters.

5. *SAS'-SS'-Ω, unbalanced and asymmetric*. This model incorporated the two variations of models 3 and 4, and thus had **7** parameters.

Note that models 3-5 were used to test for imbalances or asymmetries in the update of first-order transition probabilities and/or of Ω, which could contribute to biased perceptions of controllability (e.g. illusion of control), independent from $_{\Omega}$.

4.7 Additional model space using a task-set logics

The model space detailed above is highly generalizable because it does not make any assumption regarding the space of possible state-action pairs and it does not require that participants

know in advance the space of possible rules governing state transitions.

However, in our experiments, participants were explicitly presented with the four possible transition rules and they had the opportunity to practice enough so as to solve the task by updating — on each trial — the reliability of four possible task-sets corresponding to the four possible rules. Task-sets can be defined as "abstract constructs that signify appropriate stimulus-response grouping in a given context" [3, 4]. In our experiment, the four different task-sets can thus be represented as four fixed transition matrices linking state-action pairs to upcoming states, $TS_i(S'|S,A)$. For example, under rule C1, $TS_1(S' = square|S = circle, A = yellow) = 1$ while $TS_1(S' = square|S = circle, A = blue) = 0$. This means that, on each trial, each task set will be associated with a prediction error taking the value of 0 or 1, depending on whether the transition was compatible with the rule represented by that task-set.

$$\delta_{TS_i} = 1 - TS_1(s'|s,a)$$

Using a simple delta-rule, it is thus possible to monitor the recent amount of prediction errors associated with any given task-set. In what follow, we will refer to this key quantity as the *reliability* of the task-set and we will represent it with the symbol ρ. Its update rule is the following:

$$\rho_{TS_i} \leftarrow \rho_{TS_i} + \alpha_\rho \delta_{TS_i}$$

The advantage of tracking simultaneously the reliability of each task-set is the possibility to arbitrate amongst them when a prediction should be made about future states. In what follows, we describe two arbitration logics. The first one is called *flat* because it puts the four possible task-sets in direct competition for the control of prediction. The second one is called $\Omega - weighted$ because it is hierarchical and proceeds in two steps: first, it arbitrates between the two rules of a given type (i.e controllable or not), second, it computes the weighted reliability of each rule category. Third, it arbitrates amongst these two rule categories to determine their

relative contribution to predictions.

4.7.1 Flat arbitration

In the case of flat arbitration, the relative weight of each task-set is obtained through a simple softmax:

$$w(TS = i) = \frac{\exp(\beta_{flat}\rho_{TSi})}{\sum_{j=1}^{j=4}\exp(\beta_{flat}\rho_{TSi})}$$

4.7.2 Hierarchical Ω-weighted arbitration

Here, the relative weight of each task-sets is computed within each rule category. Considering that j=1 or j=2 index the two task-sets representing the uncontrollable rules, whereas j=3 and j=4 index the two task-sets representing the controllable rules:

$$w_{TSi} = \begin{cases} \frac{\exp(\beta_C\rho_{TS3})}{\sum_{j=1}^{j=2}\exp(\beta_C\rho_{TSi})}, & \text{if j} \leq 2 \\ \text{‘} \\ \frac{\exp(\beta_U\rho_{TS3})}{\sum_{j=3}^{j=4}\exp(\beta_U\rho_{TSi})}, & \text{if j} \geq 3 \end{cases}$$

Where β_U and β_C are the slope parameters determining to which extent the task-set with higher reliability was given a stronger weight. For $\beta >> 0$, the algorithm strongly favors the most reliable task-set even if reliability difference is small.

On each trial, a simple delta-rule is used to monitor the expected difference in the weighted prediction errors generated by the rules of each category. Due to its tight homology with the Ω variable of the main model space, we named that variable similarly.

$$\delta_\Omega = \sum_{i=1}^{i=2} w_{TSi}(1 - TS_i(s'|s,a)) - \sum_{i=3}^{i=4} w_{TSi}(1 - TS_i(s'|s,a))$$

As in the main model space, Ω is updated using the following delta-rule:

$$\Omega \leftarrow \Omega + \alpha_\Omega(P(s'|s,a) - P(s'|s) - \Omega)$$

As in the main model space, Ω is then used to compute the arbitrator ω which reflects a possibly biased inference of controllability:

$$\omega = \frac{1}{1+\exp(-\beta_\Omega(\Omega - threshold_\Omega)}$$

This arbitrator can then be used to adjust the relative weight of each task-set in the following way, so as to take into account controllability estimation:

$$w_{TSi} = \begin{cases} w_{TSi} \leftarrow w_{TSi}(1-\omega), & \text{if j} \leq 2 \\ w_{TSi} \leftarrow w_{TSi}\omega, & \text{if j} \geq 3 \end{cases}$$

Finally, the predictions of each task-set are mixed according to their weights, so that

$$p(S' = i) = \sum_{i=1}^{i=4} w_{TSi} TS_i(s'|s,a)$$

As in the main model space, the probability that the participant predicts the next state would be i (e.g. a square state) when confronted to the hypothetical state-action pair S,A (e.g. circle state, blue action) is then given by:

$$p(prediction = i) = \frac{\exp(\beta_{choice} p(S' = i)}{\sum_{j=1}^{j=3} \exp(\beta_{choice} p(S' = i))}$$

4.7.3 Variants using a task-set logics

As for the main model space, we compared different variants implementing this task-set logics, with or without Ω-weighted arbitration. Here after we describe the 10 variants corresponding to the model comparison results reported in Fig. 3. Variants implementing a flat arbitration have only an inverse temperature β_{choice} at the decision stage. Variants implementing an Ω-weighted arbitration further include two parameters controlling the inference of controllability based on Ω (i.e. $thres_\Omega$ and β_Ω).

1. **Flat arbitration, symmetric updating updating of reliabilities**: the learning rates governing the update of ρ_{TSi} are equal for any i (i.e controllable or uncontrollable task-sets. 3 parameters.

2. **Flat arbitration, asymmetric updating of reliabilities**: the learning rates governing the update of ρ_{TSi} are different for $i \in [1,2]$ (uncontrollable TS) and $i \in [3,4]$ (controllable TS). 3 parameters.

3. **Ω-weighted arbitration, symmetric updating of reliabilities, symmetric intra-dimentional arbitrations, symmetric update of** Ω: the learning rates governing the update of ρ_{TSi} and the coefficients (β) controlling the intra-dimensional arbitrations are equal for any i. The learning rate controlling the update of Ω is identical for upward and downward changes. 6 parameters.

4. **Ω-weighted arbitration, symmetric updating of reliabilities, symmetric intra-dimentional arbitrations, asymmetric update of** Ω: same as 3, with different learning rates for upward and downward changes in Ω. 7 parameters.

5. **Ω-weighted arbitration, symmetric updating of reliabilities, asymmetric intra-dimentional arbitrations, symmetric update of** Ω: same as 3, with different coefficients (β_U and β_C) controlling the intra-dimensional arbitration of controllable or uncontrollable TS. 7 parameters.

6. **Ω-weighted arbitration, symmetric updating of reliabilities, asymmetric intra-dimentional arbitrations, asymmetric update of** Ω: same as 5, with different learning rates for upward and downward changes in Ω. 8 parameters.

7. **Ω-weighted arbitration, asymmetric updating of reliabilities, symmetric intra-dimentional arbitrations, symmetric update of** Ω: same as 3, with different learning rates for the updating of reliabilities related to uncontrollable or controllable TS. 7 parameters.

8. **Ω-weighted arbitration, asymmetric updating of reliabilities, symmetric intra-dimentional**

arbitrations, asymmetric update of Ω: same as 7, with different learning rates for upward and downward changes in Ω. 8 parameters.

9. **Ω-weighted arbitration, asymmetric updating of reliabilities, asymmetric intra-dimentional arbitrations, symmetric update of** Ω: same as 7, with different learning rates for the updating of reliabilities related to uncontrollable or controllable TS. 8 parameters.

10. **Ω-weighted arbitration, asymmetric updating of reliabilities, asymmetric intra-dimentional arbitrations, asymmetric update of** Ω: same as 9, with different learning rates for upward and downward changes in Ω. 9 parameters.

4.8 Model fitting procedure

Model fitting was performed using a Variational Bayesian (VB) estimation procedure using the well-validated VBA toolbox [5]. Compared to non Bayesian methods, this approach has the key advantage of accounting for the uncertainty related to model parameters and hidden states, as well as of informing the optimization algorithm about prior distributions of parameters' values. To the exception of β_{choice}, parameters were transformed so as to restrict their variation to meaningful intervals. Thus, all learning rates where passed through a sigmoïd function ($\frac{1}{1+e^{-x}}$) limiting their variation to the [0,1] interval. Similarly, the $_\Omega$ parameter was passed through a scaled sigmoid ($-1+\frac{2}{1+e^{-x}}$) so as to limit its variation to the [-1,1] interval. Finally, the β_Ω parameter was constrained to be positive using an exponential transformation. The same transformations were used for the second model space, $beta_U$ and $beta_C$ being treated as β_Ω.

For the behavioral experiments, the prior distributions of the various learning rates and threshold parameter were innately defined as Gaussian distributions of mean 0 and variance 3, which approximates the uniform distribution over the interval of interest after sigmoid transformation. The prior distributions of β_{choice} and β_ω parameters were defined as Gaussian distribution

of mean 0 and variance 10. For the fMRI and the stress experiments, the prior distributions of every parameter was defined using the posterior mean and variance obtained from the 50 participants who passed the behavioral experiment.

Hidden states (i.e values) corresponding to transition probabilities were systematically initialized at 1/3 (equiprobability prior), while Ω was initialized at 0. The VB algorithm was not allowed to update the initial values for hidden states. Contrary to the behavioral experiment, the SS'-SAS' task was split in 4 blocks of equivalent length in the stress and fMRI experiments. Thus, we reinitialized all hidden states at their prior values at the beginning of each block to account for this discontinuity.

4.9 Simulations: model and parameter recovery

In order to ascertain that our task could discriminate participants using the SAS' from those using the best fitting SS'-SAS'-Ω scheme, we simulated 500 participants based on each scheme. For each simulated dataset, parameters were randomly drawn from Gaussian distributions whose means and variances were equal to those observed empirically in the fMRI experiment. Both models were then fitted on the two surrogate datasets of 500 simulated participants using flat priors (mean 0 and variance 3 for all parameters). We then performed one Bayesian group comparisons per dataset, in order to obtain the model selection frequencies of each model.

The quality of parameter recovery for the best-fitting model was estimated by the correlation matrix between the parameters used to generate the simulated data and the recovered parameters.

4.10 Model comparisons

Model comparisons were performed using both fixed-effects and random-effects approaches. Fixed-effects analysis assumes that only one model is used by the whole population and thus

sum the information criteria over all participants. For the sake of simplicity, in the main text and figures, we report only the results of the random-effects analyses, which treated model attribution as a random factor potentially model specific using the Bayesian group comparison [6] algorithm of the VBA toolbox.

5 Functional Magnetic Resonance Imaging (fMRI)

5.1 fMRI: acquisition

All images were collected using a 3T Siemens Magnetom Prismafit MRI scanner (Erlangen, Germany) with a 32-channel head coil. A T2*-weighted multiband echo planar imaging sequence with acceleration factor 8 (MB8) was used to acquire BOLD-fMRI whole-brain covered images (TR = 700 ms, TE = 39 ms, flip angle = 52, voxel size = 2.4 × 2.4 × 2.4 mm3 , slice gap = 0 mm, and FOV = 210 mm). This state-of-the-art sequencing protocol was optimized from the recommended imaging guidelines of the Human Connectome Project, with the fast acquisition speed facilitating the detection and removal of non-neuronal contributions to BOLD changes (http://protocols.humanconnectome.org/HCP/3T/imaging-protocols.html). The experiment was divided in 4 blocks lasting on average 7.7+/-2.1 minutes (662+/-179 volumes). We recorded participants' heartbeats using the scanner's built-in photoplethysmograph, placed on the right index finger. Respiration was measured with a pneumatic belt positioned at the level of the abdomen. Anatomical images were acquired using a T1-weigted MPRAGE sequence, using a GRAPPA acceleration factor of 2 (TR = 2300ms, TE = 3.03 ms, voxel size = 1x1x1mm, 192 transversal slices, 8° flip angle). Field magnitude and phase maps were also acquired.

5.2 fMRI: preprocessing

fMRI data processing and statistical analyses were performed using statistical parametric mapping (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK). For each session,

the first 4 volumes were automatically discarded by the scanner. Functional images were slice-time corrected, unwarped using the field maps and realigned to the mean functional image using a rigid-body registration. Functional images were then coregistered to the anatomical T1. Next, the anatomical image were segmented based on tissue prior probability maps for spatial normalisation employing DARTEL [7] and the resulting normalization matrix was applied to all functional images. Finally, all images were spatially smoothed with a 6mm Gaussian kernel, except in the decoding analysis for which unsmoothed images were used.

5.3 fMRI: mass univariate analysis

Statistical analyses of fMRI signals were performed using a conventional two-levels random-effects approach in SPM12. All general linear models (GLM) described below included the 6 unconvolved motion parameters from the realignment step. We also included the eigenvariate of signals from cerebrospinal fluid (CSF) in our GLM (fourth and lateral ventricular). Moreover, we used a retrospective image correction (RETROICOR) method to regress out physiological noise, using 10 cardiac phase regressors and 10 respiratory phase regressors obtained by expanding cosines and sines of each signal phases to the 5th order. We also included time shifted cardiac rates (lag: +6, +10 and +12s) and respiratory volume (-1 ans +5s) as nuisance regressors.

All regressors of interest were convolved with the canonical hemodynamic response function (HRF). All GLM models included a high-pass filter to remove low-frequency artifacts from the data (cut-off = 96s) as well as a run-specific intercept. Temporal autocorrelation was modeled using an AR(1) process. All motor responses recorded were modeled using a zero-duration Dirac function. We used standard voxel-wise threshold to generate SPM maps ($p<0.001$ uncorrected), unless notified otherwise. All statistical inferences based on whole-brain analyses satisfied the standard multiple comparison threshold ($p(FWE)<0.05$) at the cluster level unless notified otherwise. Prediction error and other parametric regressors were systematically

z-scored to exclude scaling effects. Reaction time regressors were log-transformed before z-scoring.

All GLM models included separate onset regressors for motor responses, for prediction trials and for the first trial of each exploratory sequence (where no prediction error was elicited). All models also included parametric regressors for reaction time and ω (reflecting controllability estimates) on prediction trials. Hereafter, we describe the 6 different variants which were used to generate the results reported in this study:

1. *Binarized PE model*. In this model, we split exploratory trials in four categories based on the prediction errors elicited by the spectator and actor models: trials where both PEs were above their median, trials where both PEs were below their median, trials where only $\delta_{sas'}$ was above its median and trials where only $\delta_{ss'}$ was above its median. This procedure was chosen to circumvent collinearity concerns in the estimation of BOLD responses, as the two prediction error terms are by definition positively correlated. Note that PE effects using conventional parametric regressions were systematically double-checked using the mixed-model approach detailed in the next subsection.

2. *PE difference model*. In this model, exploratory trials were modeled using one regressor, on top of which reaction times and the difference term $\delta_{ss'} - \delta_{sas'}$ were added as parametric regressors.

3. *PE difference model + interaction*. This model was identical to the PE difference model, except that an interaction term $zscore(\delta_{ss'} - \delta_{sas'}) * zscore(RT)$ was added to investigate whether neural RT effects interacted with the difference term.

4. *Binarized PE model with RT*. This model was similar to the Binarized PE model, except that parametric reaction time regressors were added to each of the four trial types. This model was only used to decompose the interaction term mentioned above.

5. *Controllability model*. This model was similar to the PE difference model, except that the difference term was replaced by δ_{Ω}, reflecting controllability update. Because the two controllable rules differed in the frequency of repetition of the same state (very rare in rule C1, common in rule C2), we further included a state repeat regressor, taking the value of 1 whenever a state repeat occurred and 0 otherwise. We also included $abs(\delta_{\Omega})$ in order to control for the overall amount of change in controllability estimates.

6. *Decoding*. For decoding analyses, we modeled each "miniblock" of exploration trials using a separate regressor. Since trials were pooled together, reaction times were included as a parametric regressor on motor responses in this model. Note that this model used unsmoothed functional images in order to retain a maximum of information in the spatial activation patterns. The searchlight radius was set to 12 millimeters.

5.4 fMRI: mixed-effects ROI analysis

In order to verify the robustness of our whole-brain results and inspect the time course of our parametric effects of interest, we performed mixed-effects analyses on BOLD signal filtered and adjusted for nuisance regressors. This adjusted signal was extracted from the functional clusters uncovered by whole-brain analyses and segmented into trial epochs from -3 to +16 seconds around the onset of each exploration trial (excluding the first of each streak). We then estimated the effect of each regressor of interest, at each time point, for all subjects simultaneously. Subject identity was included as a random effect and a subject-specific intercept was included. Parametric regressors were z-scored in the same way as in the mass univariate analyses. Importantly, this approach was not used for statistical inference — since doing so would constitute double-dipping — but merely for visualization purposes.

5.5 fMRI: decoding analyses

Decoding analyses were performed using the TDT toolbox [8]. Each mini-block of 6 exploratory trials was arbitrarily coded as +1 (controllable) or -1 (uncontrollable) based on the rule governing transitions. We used a leave-one-run out cross-validation scheme with 100 permutations per subject, so that classes remained balanced for training. Training was performed on the beta values associated with each miniblock (see previous section) using a Support Vector Machine (SVM) classifier (L2-loss function, cost parameter set to 1, Liblinear, version 1.94), without feature selection or feature transformation. Since we did not constrain the testing sets to have balanced classes, balanced accuracies were used when reporting the results of the searchlight analysis (12mm sphere) at the whole-brain level.

6 Supplemental Text

Estimating environmental controllability requires that the agent explores the space of possible actions as randomly as possible. Conversely, if an agent were to always select the same action A in a given state S, the empirical transition probabilities to S', $p(S'|S,A)$ and $p(S'|S)$ would always be equal, independently of the actual logical structure of the environment. From an informational viewpoint, this implies that the agent is *acausal* because S is sufficient to predict S' for an external observer. As a result, controllable and uncontrollable environments are indistinguishable both for the agent and an external observer. More formally, it can be easily proven that $H(S'|S) - H(S'|S,A) > 0$ implies $H(A|S) > 0$:

$$H(S'|S) - H(S'|S,A) > 0$$

By applying Bayes' rule to both conditional entropy terms, we obtain:

$$H(S|S') - H(S) - H(S,A|S') + H(S,A) > 0$$

By applying the chain rule of conditional entropy, we obtain:

$$H(A|S) - H(A|S',S) > 0$$

$$H(A|S) > H(A|S',S)$$

Since the entropy is non-negative by definition, $H(A|S',S) \geq 0$, therefore $H(A|S) > 0$. This relationship implies that some degree of exploration is required to obtain a non-zero transfer entropy (i.e be causal from an information theory perspective) and therefore to estimate controllability (see also, Fig. S1C).

Importantly, the causality measured by transfer entropy is entirely dependent upon the definition chosen to characterize the states and actions S, A and S'. In this study, the states S and S' were operationally defined as the successive geometrical shapes appearing on the screen, whereas the actions A were defined as the colors selected by participants. This is obviously an abstraction enabling scientific inquiry. The very existence of agents separated from their environment is highly questionable from an epistemological point of view. Instead, the operational meaning of $H(A|S) > 0$ is as simple as "different colors were selected by the participants in front of a given geometrical shape".

Another assumption of the transfer entropy framework is that the decisions of our participants were only susceptible to being conditioned by the geometrical shape displayed on the screen, that is, S. Therefore, the fact that transfer entropy increased above zero when participants explored the different options available in controllable contexts should obviously not be equated with the existence of an absolute causality emanating from the agent.

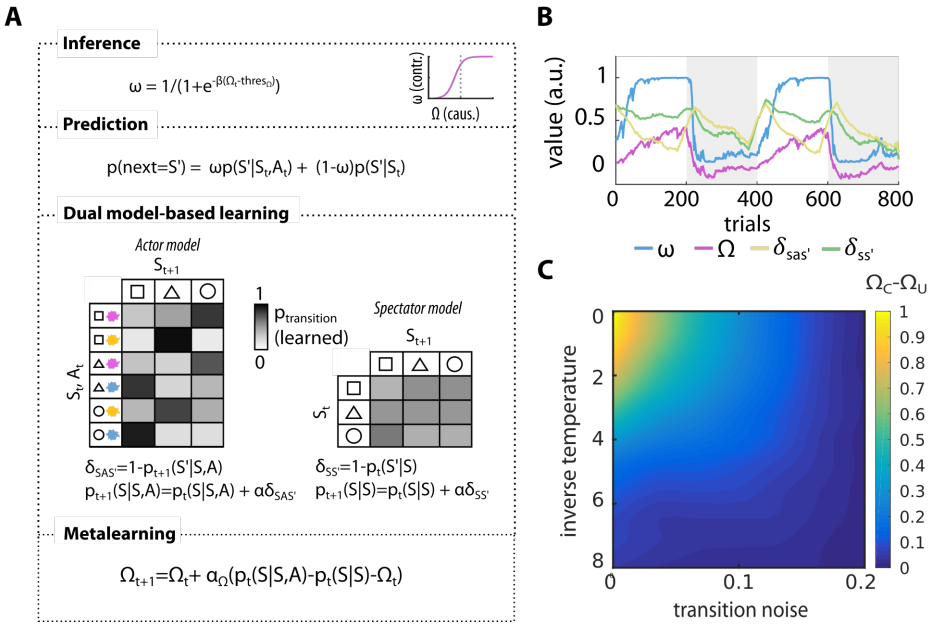

Figure S1. The SS'-SAS'-Ω architecture. Related to Figures 1 and 3.

(A) Visual representation. The SS'-SAS'- architecture consists of two parallel probability learning processes. The "actor" model updates on each exploratory trial the transition probabilities from states S and actions A to next states S'. The "spectator" model updates on each exploratory trial the transition probabilities from state S and to next state S': it therefore represents transition probabilities marginalized over actions. These two first-order processes inform a second-order process tracking the expected difference in the probability of observing next states S' given A and S, versus S only. This second order process simply uses a delta-rule to update the variable Ω which reflects the causal influence of one's own action over state transitions. Inferences about controllability are implemented by applying a threshold on Ω determining the value above which it is perceived as evidence that the environment is controllable. This thresholded Ω is further passed into a sigmoid whose slope parameter determines to which extent an Ω below or above the threshold leads to the use of the spectator or the actor model, respectively. Indeed, the resulting arbitrator variable termed ω is then used as a weight parameter controlling the mixing of transition probabilities at the time of predictions.

(B) Time courses of first order prediction errors, Ω and ωin a simulation of the task (noise level=0.05, learning rates=0.1, $thres_\Omega = -0.05$ and $\beta_\Omega = 20$).

(C) Discrimination of task conditions (C-U) by Ω in simulations varying transition noise (0 to 0.2) and the inverse temperature parameter determining to which extent the agents choose based on state values (in this simulation, the 3 states gave rewards of 0, 0.5 and 1 units, and the agents learned these values using a simple Rescorla-Wagner rule).

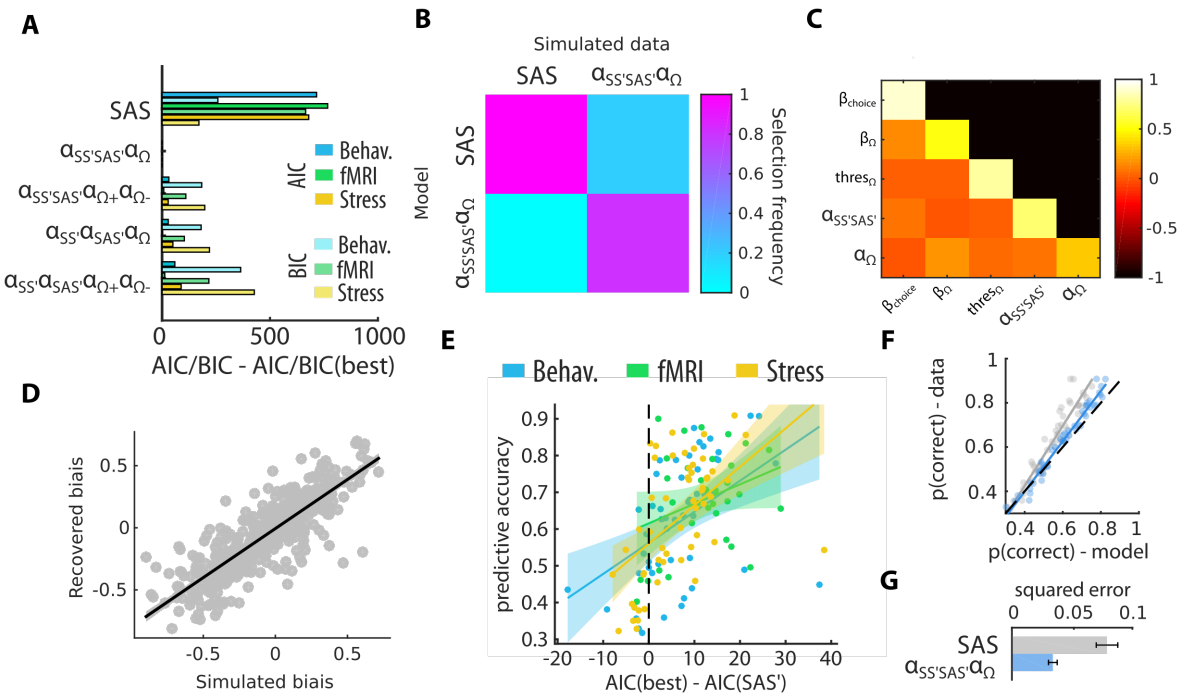

Figure S2. The advantage of the SS'-SAS'-Ω over standard SAS' model-based learning. Related to Figures 3A-C.

(A) Model comparison treating model attribution as a fixed-effect, using both the AIC and BIC. This analysis split by experiment shows that the simplest variant of the SS'-SAS'-Ω architecture (i.e. balanced and symmetrical) outperformed all other candidates and more particularly a standard SA-S' model.

(B) Model recovery analysis showed that the task could reliably discriminate the SAS' model alone from the best-fitting SS'-SAS'-Ω scheme (see Methods). The SAS' model was always identified as the most likely model when it was used to generate the surrogate data, using either BIC or AIC as a goodness of fit metric. When the SS'-SAS'-Ω scheme was used to generate the data, it was identified as the most likely model 80% of the time when using BIC (as reported in the panel) and 100% of the time when using AIC.

(C) All parameters of the best-fitting SS'-SAS'-Ω model were recovered correctly (all r>0.5 in the diagonal, except for the learning rate controlling the evolution of Ω for which r=0.31). Parameters were not correlated with each other (off-diagonal correlation coefficients between -0.02 and 0.17).

(D) In particular, the threshold parameter (thresΩ) used to analyze interindividual differences in controllability estimation and to evaluate the effect of the stress induction was highly recoverable (r=0.79).

(E) The relative advantage of this best fitting model relative to the SA-S' model was positively

correlated with the mean predictive accuracy across participants in each of the 3 experiments, which demonstrates the benefits of using such strategy.
(F) The predictive accuracy of simulated choices was tightly related to the actual accuracy of participants when using the SS'-SAS'-Ω architecture (blue). By contrast, simulated choices from the SA-S' model (gray) did not account well for the accuracy of the best participants.
(G) The absolute error between predicted model accuracies and actual accuracies of participants was greatly reduced by the best-fitting SS'-SAS'-Ω scheme as compared to the SA-S' model alone ($t(49)=6.57$, $p<0.001$).
Error bars and shaded areas represent SEM.

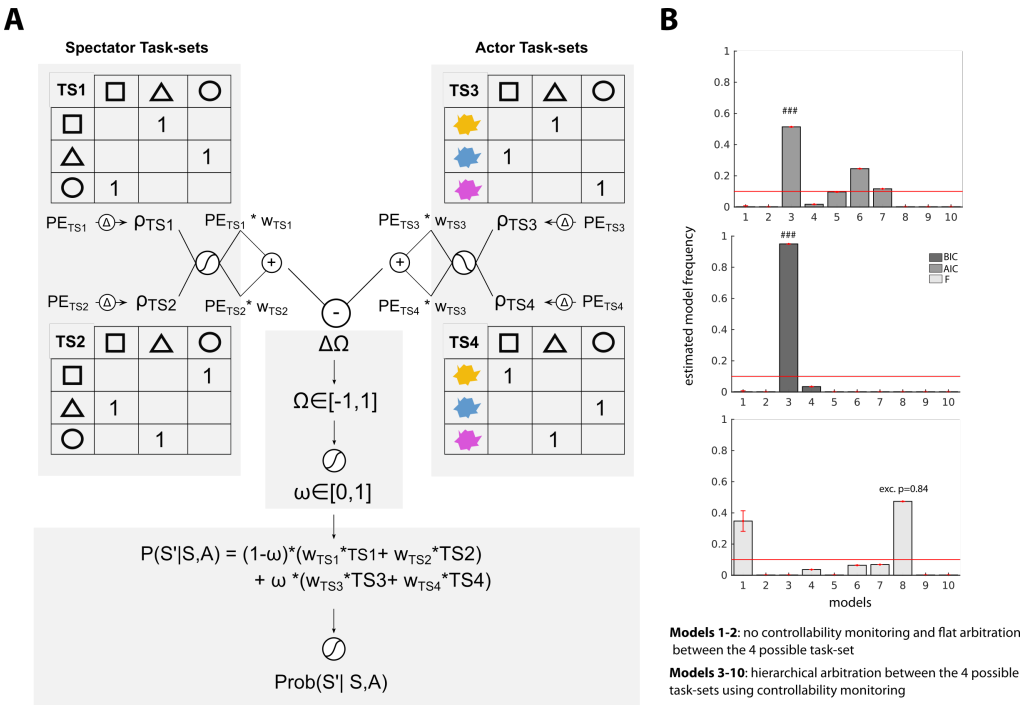

Figure S3. Generalization to a second model space using a task-set logics. Related to Figures 3A-C.

(A) Synthetic overview of the task-set scheme using Ω-arbitration. In each trial, the four possible task-sets, corresponding to the 4 different rules, generated a prediction error of 0 or 1 depending on whether the observed transition was respectively compatible or not with their corresponding rule. These prediction errors were monitored using a delta-rule, hence reflecting the reliability of each of the four task-sets given the recent history of transitions. In the Ω-arbitration scheme, an intradimensional arbitration was first used to determine the relative weights of TS1 versus TS2 (spectator rules) and TS3 versus TS4 (actor rules). These weights were in turn used to compute the weighted prediction errors associated with each task-set dimension. As for the main SS'-SAS'-Ω scheme, the difference of these weighted predictions errors were used to update Ω, which therefore reflected the relative advantage of actor task-sets over spectator task-sets. Still as in the main model space, Ω was passed through a 2-parameter sigmoïd function so as to obtain the arbitrator variable ω which was used to weight the prediction of each task-dimensions. As a result, the final weight of each task-set for prediction depended both on an intradimensional arbitration (e.g. wTS1) and on an interdimensional arbitration (ω). Oppositely, in the flat arbitration scheme used as a reference model in model comparisons, the algorithm arbitrated amongst the four possible task-sets in a single step. This latter scheme was therefore blind to variations in controllability.

(B) Bayesian model comparisons showing that the simplest -arbitration scheme (model 3) overperformed all other variants when using AIC or BIC. When using Free Energy, a slightly more complex model was selected (model 8, which used two different learning rates to update upwards or downward). The 10 different variants are described in the Methods (see subsection 4.7.3, Variants using a task-set logics).

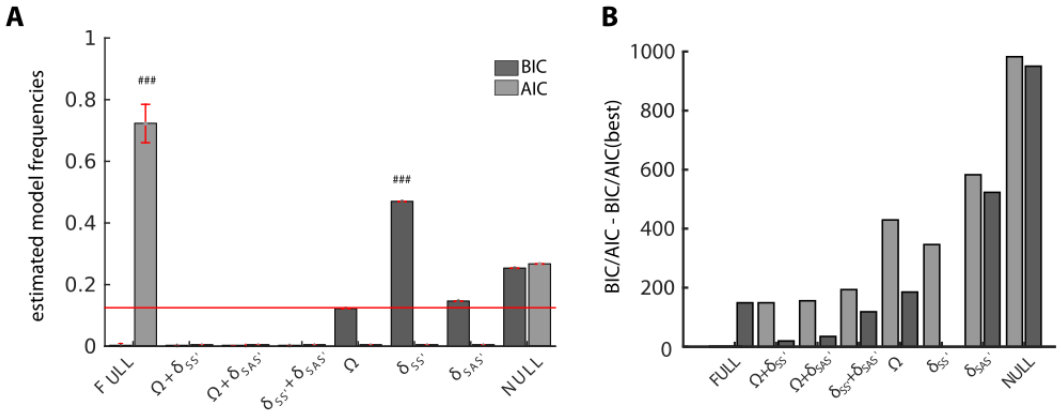

Figure S4. Model comparison related to the analysis of reaction times. Related to Figure 3D-E.

(A) Model comparison treating model attribution as a random effect. When using AIC, the most likely model was the full model which included Ω, δSAS' and δSS' as regressors. When using BIC, the most likely model included only δSS'. This inconsistency derived from the fact that the complexity implemented by the BIC is dependent on the number of data points N by the number of parameters k (i.e k*ln(N)), while the AIC does not (2*k). Given the very large number of exploratory trials (ie. N=562+/-162), this penalization was more stringent for the BIC. Moreover, it should be emphasized that the main goal of this reaction time analysis was to establish the dissociation between the actor and spectator model. Since the existence of the actor model is already well-established by a large literature on model-based reinforcement learning [1, 2, 9], the fact that the best fitting model based on the BIC criterion included δSS' may already been seen as support for the dissociation hypothesis.

(B) The same conclusions held when treating model attribution as a fixed effect.

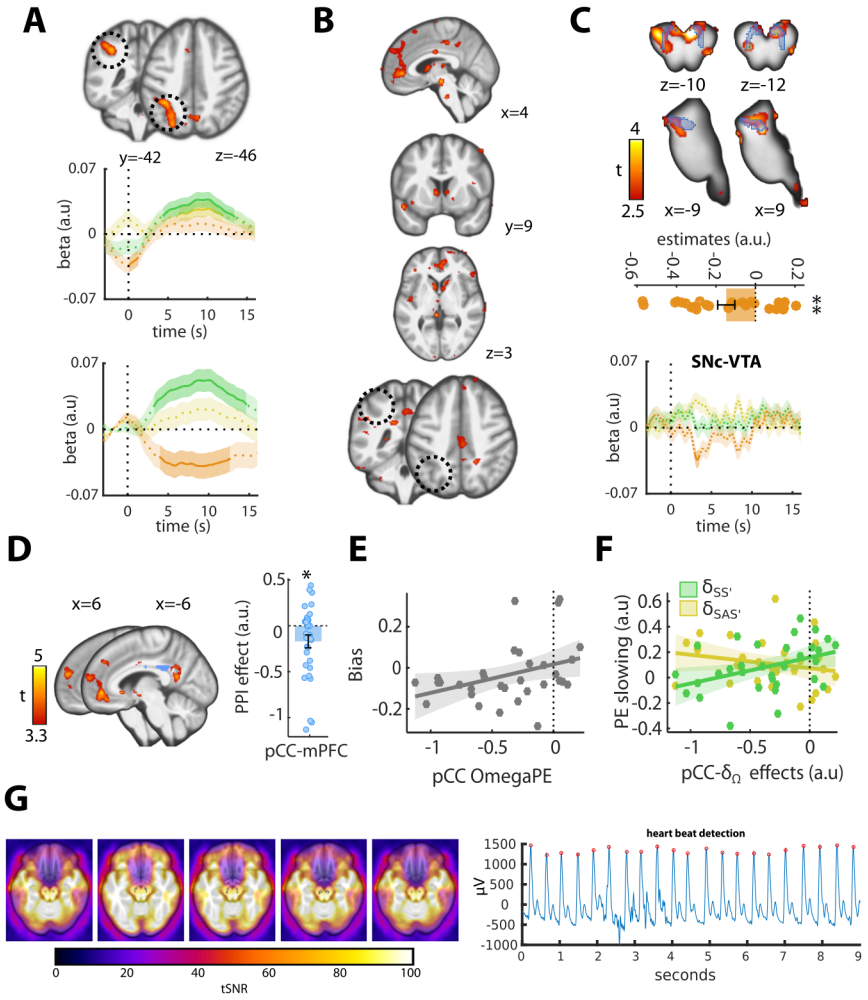

Figure S5. Supplementary neuroimaging analyses and data quality. Related to Figure 4. (A) The left intraparietal sulcus (IPS) was significant when contrasting trials where δSAS' was above the median while δSS' was below (top). However, this region also responded when contrasting trials where both PEs were above their median to trials where they were both below (Fig. 4A). Our mixed effect analysis reflected this latter effect but did not clearly support a dissociation of the two prediction errors, nor an encoding of the difference term δSAS'-δSS' (middle) in the IPS. The expected pattern was only visible when performing a baseline correction (bottom), hence casting doubt on the robustness of the results.
(B) This led us to perform an additional analysis in which we contrasted the parametric effects of δSAS' and δSS' obtained from two separate GLMs. This analysis confirmed the differential encoding of each type of prediction error in the mPFC and the nucleus accumbens (top three

maps, voxel-wise threshold: p<0.005 uncorrected, cluster-wise: p< 0.05^{FWE}), as well as in the dopaminergic brainstem (see next, c). However, the left IPS did not appear in this contrast (bottom maps), which led us to reject the hypothesis that the two prediction errors dissociate in this structure.
(C) Given that the nucleus accumbens and the mPFC are densely irrigated by dopamine which plays a key role in prediction error signaling, we further evaluated BOLD responses in this structure. Indeed, a functional cluster overlapping with the VTA could be seen in the parametric maps testing for the difference term δSAS'-δSS' and in the analysis performed in (b). Thus, we analyzed brain stem responses of the 26 participants for which respiration and heart rate signals could be used to denoise recordings and improve signal to noise ratio in the VTA (see panel g). Parametric maps (small volume correction: p< 0.05^{FWE}) and a ROI analysis within an anatomical mask [10] covering the substantia nigra that the ventral tegmental area [delimited in blue, t(25)=3.31, p=0.002] confirmed that dopaminergic brain stem encoded the tested for the difference term δSAS'-δSS'. The same conclusion was also reached when using a mixed-effects analysis.
(D) A psychophysiological interaction (gPPI) analysis revealed the existence of a controllability-dependent coupling between the mPFC and the pCC (voxel-wise threshold: p<0.001 uncorrected, cluster-wise threshold: p< 0.05^{FWE}).
(E) Update signals $\delta\Omega$ in the pCC correlated across participants with the threshold parameter [r=0.42, p=0.016], so that participants with more salient encoding of $\delta\Omega$ needed less causal evidence to infer that the rule was controllable.
(F) Consistently, in participants with stronger (i.e more negative) controllability update signals in the pCC, the slowing of decision times was increased in response to prediction errors generated by the actor model and reduced in response to those generated by the spectator model [δSS': r=0.40, p=0.02, δSAS': r=-0.22, p=0.148; difference of rSS' and rSAS': z=2.07, p=0.02]. Deactivations of the pCC might thus promote reliance on the "actor" model by integrating afferent mPFC signals, which themselves encode the relative uncertainty of S-S' versus SA-S' transitions.
(G) The mean tSNR of smoothed normalized data in the dopaminergic brainstem mask (delimited in black at the center of each slice) was comparable to that of ventral cortical structures (mean value of the group: 78.9). Recorded physiological signals allowed reliable detection of individual heart beat and respiration signals, which were used to form retrospective correction of physiological motion (RETROICOR).
Error bars and shaded areas represent SEM.

	Rule (main effect)			**Time (main effect)**			**Rule x Time (interaction)**		
	df	F	p	df	F	p	df	F	p
behavior	1 / 47	9.58	0.003	5 / 235	4.81	<0.001	5 / 235	2.05	0.07
fMRI	1 / 31	1.73	0.2	7 / 217	1.94	0.06	7 / 217	1.73	0.1
stress	1 / 53	8.29	0.006	7 / 371	2.56	0.006	7 / 371	2.05	0.048

Table S1. Related to Figure 2A. Prediction accuracies of participants in the 3 experiments analyzed using a 2-way repeated-measures ANOVA.

IBASPM116	BA	side	cFWE	cFDR	extent	pFWE	pT	X	Y	Z
Angular	40	R	<0.001	0.000	125	0.592	4.85	36	-49	38
Parietal Sup	7	L	<0.001	0.000	138	0.699	4.72	-27	-61	47
Insula	47	L	0.090	0.037	38	0.769	4.63	-27	29	-1
Precentral (dlPFC)	9	L	<0.001	0.000	144	0.880	4.46	-42	5	35
Supp Motor Area	6	L	0.107	0.037	36	0.984	4.15	-6	14	56

Table S2. Related to Figure 4A. Clusters responding to the contrast "Actor and spectator PE above their median">"Actor and spectator PE below their median" (minimal cluster extent: 10, voxel-wise threshold: p<0.001 uncorrected, cluster-wise threshold: $p< 0.05^{FWE}$). The asterisk denotes a cluster which only approached cluster-level correction for multiple comparison using FWE correction, but passed it based on the FDR criterion.

IBASPM116	BA	side	cFWE	cFDR	extent	pFWE	pT	X	Y	Z
δsas' > δss' (categorical)										
Cingulum Ant	10	R	0.004	0.004	76	0.91	4.4	15	50	11
δss' > δsas' (categorical)										
Parietal Sup L	7	L	<0.001	<0.001	256	0.711	4.7	-27	-67	59
δsas' > δss' (parametric)										
Brainstem	SN	L	0.519	0.380	49	0.75934	4.638	-15	-16	-10
Cingulum Ant R	32	R	<0.001	0.001	237	0.78478	4.604	3	41	-1
Caudate R*	Ca.	R	0.126	0.174	81	1	3.7572	6	2	8
	Ca.	L				1	3.46	-9	2	17
δsas'- δss' (parametric)										
Brainstem*	SN	L	0.762	0.375	12	0.134	5.59	-15	-16	-10
Cingulum Ant	32	R	0.001	0.004	85	0.454	5.02	9	41	17
Caudate*	Ca.	L	0.150	0.130	30	0.785	4.61	-6	11	5
Caudate	Ca.	R	0.044	0.064	43	0.92	4.38	6	2	8

Table S3. Related to Figure 4B. Clusters responding to the contrast "Only actor PE above its median">"Only spectator PE above its median" and vice-versa (categorical contrasts, minimal cluster extent: 10, voxel-wise threshold: p<0.001 uncorrected, cluster-wise threshold: $p< 0.05^{FWE}$). Clusters encoding differently δsas' and δss' (δsas' > δss', parametric, minimal cluster extent: 10, voxel-wise threshold: p<0.001 uncorrected, cluster-wise threshold: $p< 0.05^{FWE}$). Cluster encoding the difference between the two prediction error terms, as estimated using distinct subject-level GLMs (δsas'- δss', parametric, minimal cluster extent: 10, voxel-wise threshold: p<0.005 uncorrected, cluster-wise threshold: $p< 0.05^{FWE}$). Clusters denoted by an asterisk did not survive correction for multiple comparisons at the whole-brain level but were reported because they were significant when restricted to an a priori anatomical mask (brainstem, see Fig. S5C) or because they were significant in one analysis and close from significance in another (caudate).

IBASPM116	BA	side	cFWE	cFDR	extent	pFWE	pT	X	Y	Z
Frontal Inf Oper	13	R	0.008	0.020	64	0.199	5.45	42	11	11
SupraMarginal	40	R	0.000	0.000	183	0.235	5.36	63	-40	26
Cingulum Mid		R	0.013	0.022	58	0.705	4.70	3	-7	29
Cingulum Mid	24	R	0.028	0.036	49	0.959	4.26	6	-19	44
Parietal Inf	40	L	0.090	0.080	36	0.997	3.98	-63	-37	44

Table S4. Related to Figures 4D-E. Clusters encoding negatively the controllability prediction error term $\delta\Omega$ (minimal cluster extent: 10, voxel-wise threshold: p<0.001 uncorrected, cluster-wise threshold: $p< 0.05^{FWE}$). The asterisk denotes a cluster which only approached correction for multiple comparison but which was included to show that the left TPJ

IBASPM116	BA	side	cFWE	cFDR	extent	pFWE	pT	X	Y	Z
Frontal Mid L	8	L	0.001	0.000	318	0.022	5.88	-36	26	44
Calcarine L	17	L	0.000	0.000	380	0.095	5.25	-9	-103	-1
Parietal Inf R	40	R	0.000	0.000	691	0.131	5.10	57	-46	38
Frontal Sup Medial R	8	R	0.002	0.001	252	0.140	5.07	3	26	53
Precuneus L	7	L	0.000	0.000	573	0.146	5.05	-9	-64	47
Frontal Mid R	32	R	0.000	0.000	487	0.254	4.78	27	38	17

Table S5. Related to Figures 4F-G. Clusters in which the balanced decoding accuracy of controllability prediction error survived correction for multiple comparisons ($p< 0.05^{FWE}$, voxel-wise threshold: p<0.001 uncorrected).

	mean controllable group	mean uncontrollable group	t (df=26)	p
αSS'SAS / first-order learning rate	0.51	0.5	-0.12	0.91
αΩ / Ω learning rate	0.28	0.24	-1.46	0.15
Beta(Choice)	3.43	2.96	-0.93	0.35
Beta(Ω)	17.23	21.72	0.44	0.44
Threshold Ω	-0.06	0.03	2.82	**0.007**
Arbitrator ω (average value)	0.65	0.5	-2.64	**0.011**

Table S6. Related to Figure 5C. Best-fitting parameters estimated in the stress experiment in participants exposed to controllable or uncontrollable stressors.

Supplementary References

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning *66*, 585–595.

Lee, S., Shimojo, S., and O'Doherty, J. Neural Computations Underlying Arbitration between Model-Based and Model-free Learning *81*, 687–699.

Monsell, S., Sumner, P., and Waters, H. Task-set reconfiguration with predictable and unpredictable task switches *31*, 327–342. 00000.

Collins, A.G.E. and Frank, M.J. Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *120*, 190–229.

Daunizeau, J., Adam, V., and Rigoux, L. VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data *10*, e1003441. 00158.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. Bayesian model selection for group studies *46*, 1004–1017.

Ashburner, J. A fast diffeomorphic image registration algorithm *38*, 95–113. 05391.

Hebart, M.N., Görgen, K., and Haynes, J.D. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data *8*. 00170.

Daw, N., Gershman, S., Seymour, B., Dayan, P., and Dolan, R. Model-Based Influences on Humans' Choices and Striatal Prediction Errors *69*, 1204–1215. 00849.

Pauli, W.M., Nili, A.N., and Tyszka, J.M. A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei *5*, 180063. 00051.