Understanding human meta-control and its pathologies using deep neural networks

Kai Sandbrink, Laurence Hunt, and Christopher Summerfield

Department of Experimental Psychology
University of Oxford
Oxford, UK

Correspondence: kai.sandbrink@psy.ox.ac.uk, christopher.summerfield@psy.ox.ac.uk

Abstract

In mammals, neurons in the medial prefrontal cortex respond to action prediction errors (APEs). Here, using computational simulations with deep neural networks, we show the that this error monitoring process is crucial for inferring how controllable an environment is, and thus for estimating the value of control processes (meta-control). We trained both humans and deep reinforcement learning (RL) agents to perform a reward-guided learning task that required adaptation to changes in environmental controllability. Deep RL agents could only solve the task when designed to explicitly predict APEs, and when trained this way, they displayed signatures of meta-control that closely resembled those observed in humans. Moreover, when deep RL agents were trained to over- or under-estimate controllability, they developed behavioural pathologies matching those of humans who reported depressive, anxious or compulsive traits on transdiagnostic questionnaires. These findings open up new avenues for studying both healthy and pathological meta-control using deep neural networks.

Introduction

Humans and other animals have evolved dedicated neural processes that help them achieve their goals, known collectively as cognitive control[1–3]. Control processes help ensure that actions are realised as intended, even in the face of uncertainty, conflict, or external perturbation. For example, if playing tennis on a windy day, or when tired, control processes might be engaged to help a player hit the ball more accurately. However, the value of control depends on the action-outcome statistics of the environment[4]. Intuitively, cognitive control is most helpful when the environment is neither trivially easy nor impossibly hard to control, because in the former case automatic processes may suffice, whereas in the latter control is futile. Biological agents have evolved neural mechanisms that estimate the value of deploying control processes[5], which have been collectively called 'meta-control'. However, the computational processes that make meta-control possible, and their neural implementation, remain unclear[6,7].

Recently, neuroscientists have rediscovered connectionist (or neural network) models as tools for explaining cognition and brain processes[8–11]. Deep network models learn to master complex tasks by trial and error, as model parameters are gradually adjusted via an optimisation process[12]. Deep networks can account for diverse behavioural and neural signatures that occur as animals perform perceptual and cognitive tasks[13–16]. Here, we used networks trained with deep reinforcement learning (deep RL)[17], an emerging computational paradigm for modelling the brain[18,19] as a theoretical framework for studying of meta-control. We asked under what computational constraints a deep RL agent might learn to optimally "control itself", i.e. to engage control processes when they are demanded by the environment[7,20–22].

The engagement of cognitive control has long been linked to processes that monitor for errors. In mammals, response errors lead to heightened neural signals contiguous with action onset, that are generated in brain regions including the medial prefrontal cortex[23–27]. These have been called "action prediction errors" (APEs), because they signal a mismatch between intended and realised actions[28,29]. APEs may offer information about how controllable the world is, and thus can help regulate the subsequent deployment of control. For example, in paradigms like the Stroop or Eriksen Flanker task, or during task switching, response errors lead to more cautious responding[30,31] and lower subsequent error likelihoods[32]. In deep RL agents, learning is naturally guided by *reward* prediction errors, but networks do not typically come equipped with means to explicitly encode whether their actions are translated into their intended consequences (i.e. to compute APEs) and thus to estimate environmental controllability. Here, thus, using a reward-guided learning ('bandit') task we asked whether providing deep RL agents with an additional module that computed APEs would help it learn an optimal meta-control policy.

We report that deep RL agents can learn an approximately optimal meta-control policy, but – critically – only when equipped with explicit mechanisms for computing APEs. When trained to predict the likelihood of errors, deep RL agents show patterns of meta-control that closely resemble humans recruited to perform the same bandit task. Moreover, by introducing subtle biases in the way neural networks compute APEs, we

can simulate patterns of disrupted behaviour characteristic of human psychological disorders, such as learned helplessness and compulsivity[33]. This provides a computational framework, based in deep learning, for understanding meta-control and its pathologies.

Results

To study meta-control in humans and deep networks, we employed a reward learning task that involves choosing between two actions that pay out a reward with unknown probability (a 'bandit' task). The variant we chose, known as the "observe vs. bet task", is designed to disambiguate information-seeking from reward-harvesting decisions[34–36] (**Fig.1A**). On each trial, participants can decide to "bet" on one of the two bandits, yielding an outcome (rewarded or unrewarded, depending on the payout probability) that is not immediately disclosed. Alternatively, they can choose to "observe" whether each of the two bandits would have paid out or not, a choice that yields information but no reward. To introduce action prediction errors, we adapted this task by introducing a probability $p(\text{flip})$ that "bet" choices would be randomly reversed from chosen to unchosen bandit after participants placed their bet (**Fig. 1C**). We define the "controllability" ξ of the environment as $1 - 2 \times p(\text{flip})$ so that when $\xi = 0$ the observer has no control over the environment, whereas $\xi = 1$ is the standard case in which actions always play out as intended. Bandit payout probabilities were set to $p(r_1)$=1 and $p(r_2)$=0 for the "human" version of the task, and they reversed with a small probability on each trial (see Methods; although the payout probability for choosing the correct bandit would be the same for $p(\text{flip}) = 0$ and a reduced $p(r_1)$, our setup differs in that participants receive fully informative feedback on observe trials, and receive information about $p(\text{flip})$ on bet trials that would be unavailable otherwise). To facilitate learning, at the end of each block of 50 trials, a screen was shown depicting the overall pattern of choices and outcomes, using a previously described method[36] (**Fig.1B**).

We used this task to study whether humans and deep networks could adapt to the controllability of the environment. Human participants were randomly split into two groups that were first trained and then evaluated on different subsets of the discrete values $\xi \in \{0.125k \mid 0 \leq k \leq 8, k \in \mathbb{N}\}$ (see Methods). During test blocks, the level of controllability ξ was not explicitly cued, and had to be inferred on the fly from the pattern of response errors (during training human participants were given overt cues to assist learning; see Methods). We began by using agents' tendency to "observe" as proxy for their estimate of the controllability of the environment. Intuitively, if $\xi = 0$ then each realised action is made entirely at random (because $p(flip) = 0.5$) and each "observe" is wasteful, because it yields no useful information, but incurs an opportunity cost. By contrast, when $\xi > 0$ observation helps agents learn the value of each bandit for subsequent trials. We verified this intuition by calculating the optimal policy under each level of controllability as a partially observable Markov decision process using an approach called SARSOP[37]. This optimal model revealed that in our task, it is only advisable to "observe" for efficacy levels in excess of $\xi \sim 0.4$. Observation is also unwarranted towards the end of a block, as the opportunity cost exceeds the benefit of continuing to harvest at risk (**Fig. 1D**).

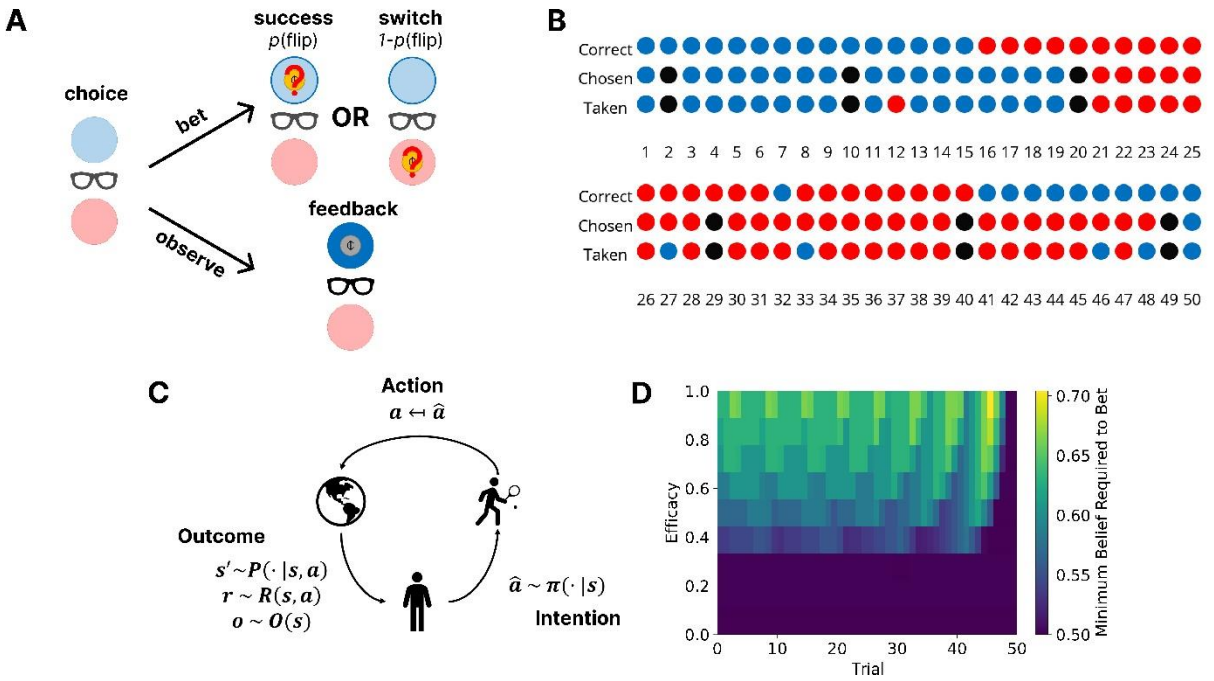

Figure 1. Task overview. (**A**) The adapted *observe or bet* task (pictures taken from implementation for human behavioural experiment). On each trial, participants have the option of choosing between betting on one of two bandits (shown here as red and blue discs; to potentially harvest reward) and observing (shown here as glasses; to gather information); on bet trials, bets are potentially switched to a bandit other than the one the participant chose. (**B**) Delayed feedback screens shown to the participants at the end of an episode provide complete information about all transitions. The top row represents which bandit paid out on which step, the middle row which one action the participant chose, and the bottom row which action was selected. Black dots indicate an observe action, and red and blue dots the corresponding bandit choices. (**C**) The process by which an agent interacts with the task can be modelled as two steps, in which an agent first marks which action they intend to take; then, an action is selected; and only then is the outcome computed (environmental transition, rewards, and observation). (**D**) The amount of evidence that is required to justify the decision to bet (cold-hot colour map), plotted as a function of the number of steps elapsed (x-axis) and the degree of efficacy (y-axis), as evaluated by SARSOP. In higher efficacy settings, more evidence is required that a given arm is correct to justify taking.

Learning meta-control: humans

We recruited n=182 human participants (n=111 after exclusions, see Methods) to take part in the study, which was pre-registered at https://osf.io/9ewt8 (data from a pilot study with very similar results is described in Supplement 3). The study took part over three sessions on separate days, with training occurring on day 1, and test (task 1) on day 2 (an additional test, called task 2 and described below, occurred on day 3). In **Fig. 2A** (left panel) we plot human observe frequencies on test blocks (from task 1) across the entire range of values of ξ. Without being explicitly cued, and starting from about trial 15, participants began to observe more frequently in more controllable environments (average observation rate: 0.144 ± 0.012, mean ± SEM, n=46 for ξ=1; 0.107 ± 0.010, n=65 for ξ=0.5; 0.070 ± 0.011, n=46 for ξ=0; $p<0.05$ for all pairwise comparisons according to Tukey's HSD; see Supplement for full statistics). As shown in **Fig. 2B** (left panel) they

were also more likely to intend correct bets when the environment was more controllable over the same range (average proportion of bets placed on correct bandit: 0.707 ± 0.028, mean ± SEM, n=46 for ξ=1; 0.566 ± 0.016, n=65 for ξ=0.5; 0.478 ± 0.012, n=46 for ξ=0; p<0.05 for all pairwise comparisons according to Tukey's HSD; see Supplement for full statistics). Collectively, these differences mean that humans achieved greater rewards in the high-controllability settings (total rewards: 29.37 ± 1.06, mean ± SEM, n=46 for ξ=1; 24.66 ± 0.58, n=65 for ξ=0.5; 22.47 ± 0.52, n=46 for ξ=0; p<0.05 for pairwise comparisons ξ=1 vs. ξ=0.5 and ξ=1 vs. ξ=0, p<0.1 for ξ=0.5 vs. ξ=0 according to Tukey's HSD; see Supplement for full statistics). Indeed, combining training and test trials, we can see that in humans both the frequency of "observe" choices, and aggregate reward, vary approximately linearly with the controllability of the environment (**Fig. 2C-D**, left panels). After each episode, participants were asked to indicate what level of environmental controllability they believed they were operating under, using a slider; their choices indicate they were able to explicitly estimate environmental controllability even on test blocks, which were not disambiguated with a cue (**Fig. S1**). Thus, humans adapted their behaviour to the controllability of the environment, similar to an ideal agent.

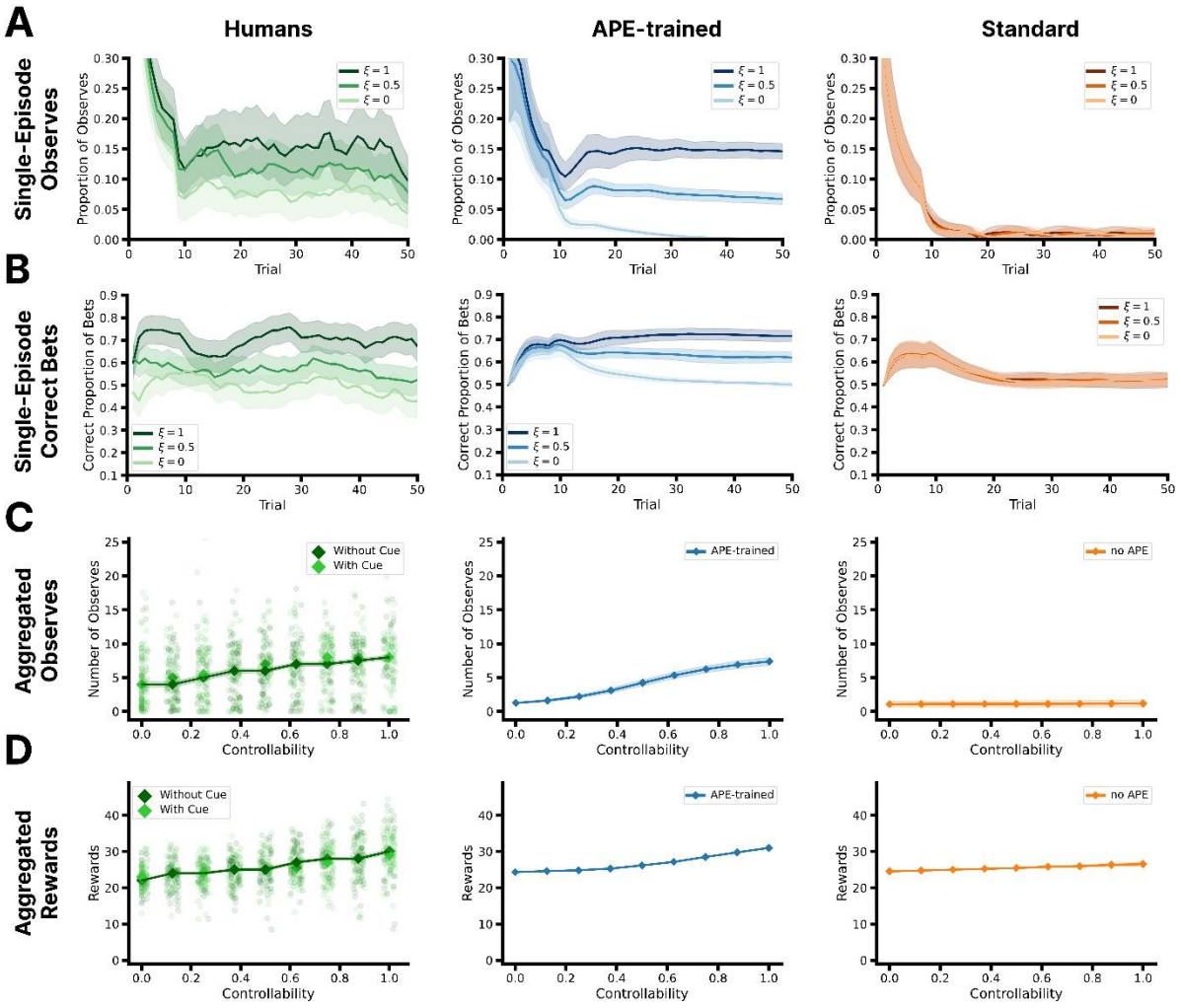

Figure 2. Humans can correctly solve the first *observe or bet* task (in a way that is well-modelled by the APE-trained networks). (**A**) Average proportion of observes over the course of each episode across (*left*) humans, (*middle*) APE-trained networks, and (*right*) no-APE networks

(see below) for three different efficacy levels, (*dark*) ξ = 1, (*medium*) ξ = 0.5, (*light*) ξ = 0. Moving-average smoothed with a window of length 8. For the networks, the average is taken over the policy probabilities. Shaded region represents standard error of the mean. (**B**) Same plot at in A, but for the probability of intending to place a bet on the right arm. (**C**) Average number of observations across different efficacy levels for (*left*) humans, (*middle*) APE-trained networks, and (*right*) no-APE networks. For humans, individual samples are shown; for the networks, the shaded regions represent standard error of the mean. (**D**) Same plot as in (C), but for the number of rewards obtained in an episode.

Learning meta-control: deep networks

Next, we studied the conditions under which deep neural networks were able to "meta-learn" how to adjust their behaviour to the controllability of the environment[38,39]. In this version of the task, we introduced probabilistic bandits, so the high-paying bandit paid out with probability $p(r_1)$=0.9, and the low-paying bandit with $p(r_2)$=0.1. Agents were trained on blocks defined by $\xi \in \left[0, \frac{1}{3}\right] \cap [\frac{2}{3}, 1]$ and evaluated on blocks with the same disjoint values of ξ. We trained deep RL agents to maximise reward using REINFORCE[40] with baselines, a machine learning method that uses deep networks to learn a mapping from observations onto optimal action probabilities (**Fig. 3A**). The network was equipped with a long short-term memory (LSTM) layer[41], which incorporates recurrence (activation memory) and gating in a way that partially resembles the prefrontal cortex[42]. This is necessary to learn within a block (i.e. to adapt policy within the block on the basis of past actions, observations and rewards, also known as "learning to reinforcement learn"[43,44]). The network was trained with backpropagation to maximise reward across several hundred training episodes.

We created three variants of the deep RL network. All variants receive as inputs the feedback received on the previous turn o_{t-1}, if any, alongside both the action realised on the previous trial a_{t-1}, and the originally intended action $\hat{a}_{t-1}$ (i.e. that which was chosen before any flip was applied). They output intended action probabilities for the current trial $p(\hat{a}_t)$, which are translated into behaviour a'_t. The *standard* variant has no further inputs or outputs, and thus has to infer the level of controllability implicitly, from frequency of match or mismatch between a and a' in its inputs. Other variants explicitly compute an APE, which is defined as

$$\delta_{APE,t} = \begin{cases} 0, & \text{if } a_t = \hat{a}_t \\ 1, & \text{otherwise} \end{cases}$$

The ξ*-input* variant receives the environmental controllability ξ directly, as an input signal, but is otherwise identical to the standard variant. The *APE-trained* network produces an additional output that estimates an error likelihood signal, i.e. the probability of an intended bet action switching, $\hat{p}_t(\text{flip})$. This scales directly with controllability of the environment ξ, which can be estimated using $\hat{\xi}_t = 1 - 2\hat{p}_t$. By training the network to predict APEs, the network learns to implement a delta rule:

$$\hat{p}_{t+1} = \hat{p}_t + \alpha \cdot \left[\hat{p}_t - \delta_{APE,t}\right]$$

This is similar to the proposed function of the medial prefrontal cortex in frameworks that emphasise the learning of error likelihoods or the expected value of control[32,45]. Relative to the other networks, the APE-trained variant has just as single additional node that encodes this prediction (**Fig. 3A**). Finally, we evaluated the performance of *single-setting* networks that had the same architecture as standard networks, but were trained exclusively on MDPs with a given fixed value of ξ (rather than learning to infer ξ anew from experience on each block).

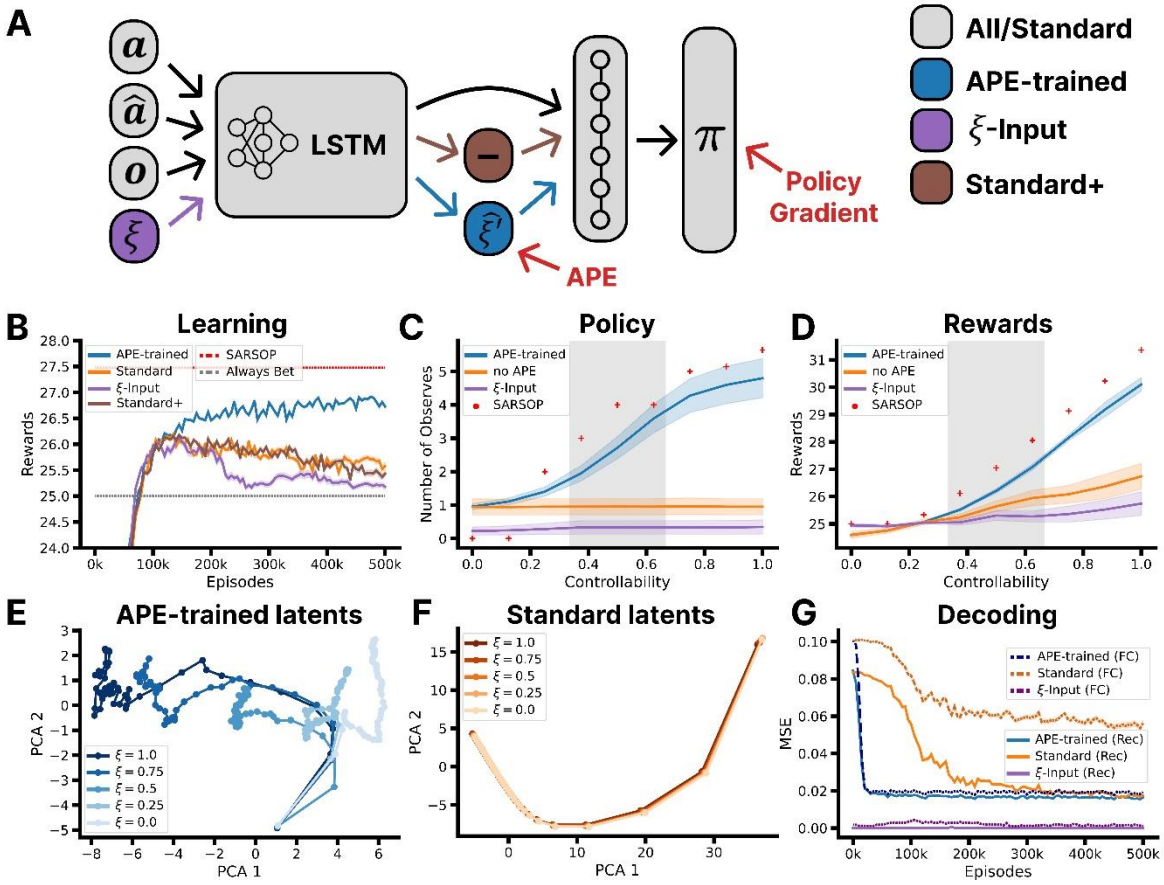

Figure 3. Training deep reinforcement learning agents on the task. (**A**) Tested network architectures showing the architecture that is shared for both APE-trained and standard networks, as well as the APE-specific efficacy readout. (**B**) Learning curves for task 1 for (*blue*) APE-trained, (*orange*) no-APE neural networks, (*green*) networks trained on individual controllability settings, (*purple*) networks that receive the ground-truth controllability $\boldsymbol{\xi}$ as an additional input, and (*brown*) no-APE neural networks with an additional hidden unit as measured by the average number of rewards achieved per episode evaluated on the test set. Upper and lower bounds on performance are given by performance achieved by (*dashed red*) SARSOP and (*dashed grey*) always betting on the same test set. (**C**) The mean number of observes over the course of an episode across all 5 model instantiations. The APE-trained models (*blue*) observe more frequently in high-efficacy episodes, whereas the no-APE models (*orange*) observe the same amount across all efficacy levels. (**D**) The mean rewards accumulated over the course of an episode across all 5 model instantiations. The APE-trained models (*blue*) earns more reward in high efficacy episodes; the difference is more slight for the (*orange*) standard models. (**E**) First two principal components for 100 trajectories across (*darkness*) five different efficacy levels for a sample APE-trained neural network. (**F**) Same as in (E), but for the no-APE networks. (**G**) Population decoding across all model instantiations from the recurrent layer (Rec) and fully

connected layer (FC) across the course of training for (*blue*) APE-trained and (*orange*) standard networks. Accuracy is measured by mean-squared-error averaged over the second half of each episode for 100 repetitions across all 10 model instantiations. In all plots, shading shows 1 S.E.M. over ten models trained from different random seeds.

We observed that the APE-trained network was able to solve task 1, converging just below the maximum reward level obtainable, as revealed by SARSOP. However, the standard and APE-input variants struggled, reaching a level of reward that was only slightly higher than that expected under random choice (**Fig. 3B**). To ensure that the *standard* network did not fail because it had one unit fewer than the APE-trained version, we also included a network called *standard+* that was parameter-matched to APE-trained; this did not improve performance (and for brevity we do not consider it further). Thus, computing APEs seems to be important for learning a meta-control policy.

Neither the *standard* nor the ξ*-input* network learned to respond to the controllability of the environment. This can be seen in **Fig. 3C**, which plots the average frequency of "observe" choices as a function of ξ. For the *standard* and ξ*-input* networks, this relationship is flat (or mostly flat). However, APE-trained networks (like humans) observe more in more controllable environments, and so obtain higher levels of reward (**Fig. 3D**). The proportion of "observe" choices over efficacy, and the average reward, for the APE-trained and standard networks are very similar to those obtained by SARSOP (**Fig. 3C-3D**), revealing that APE-trained networks (but not other variants) learn the optimal meta-control policy. The time-varying pattern of observation and reward in these networks is also strikingly similar to that seen in humans, alongside which they are plotted in **Fig. 2A-D** (centre and right panels). The APE-trained model's policy therefore predicted human choices significantly better than the standard model for all but the lowest efficacy levels (average likelihood: 0.648 ± 0.033, mean ± SEM for APE-trained and 0.519 ± 0.018, mean ± SEM for standard, n=46, for ξ=1; 0.610 ± 0.024, mean ± SEM for APE-trained and 0.481 ± 0.014, mean ± SEM, n=65 for ξ=0.5; 0.562 ± 0.022, mean ± SEM for APE-trained and 0.534 ± 0.021, mean ± SEM for standard, n=46 for ξ=0; estimated model frequencies of 0.9955 ± 0.0006 and 0.0045 ± 0.0006, mean ± SEM, n=111 and exceedance probabilities of ~1 and ~0 for APE-trained and standard, respectively; see Supplement for full statistics and Methods for details).

Neural representations of control in deep networks

Next, we studied how controllability was internally represented in the APE-trained networks and other variants. We used a dimensionality reduction method to visualise the representation of controllability in the population respond of hidden layer units for the LSTM, both for *APE-trained* and *standard* networks. In APE-trained networks, the neural response over trials is organised in the first two principal components along a controllability axis, with activity under low- controllability conditions at one end of PC1 and high- controllability conditions at the other (**Fig. 3E**). However, no such segregation is observed for the standard networks (**Fig. 3F**). For completeness, we trained a classifier to decode ξ from the recurrent layer (full lines) and the final fully connected linear layer (dashed lines) of the *APE-trained*, ξ*-input* and *standard* networks (**Fig. 3G**). Consistent with the neural state space analysis, decoding was much poorer in the recurrent layer of standard networks. Surprisingly, however, with sufficient training controllability could be

decoded from the final layer of standard networks just as readily as it could from APE-trained networks, and could be decoded throughout the ξ-input network with even greater accuracy (**Fig. 3G**). This implies that explicitly computing APE signals is not required to encode environmental controllability *per se*, but instead helps format information about controllability in a way that makes it most likely to be read out to guide behaviour. In line with this, when we use an encoding model to see what fraction of variance of individual neurons is explained by ξ, the APE-trained networks do indeed represent controllability more strongly (average encoding r^2 over final 25 trials: 5.25e-3 ± 1.67e-3, mean ± SEM for APE-trained, -5.34e-5 ± 3.38e-5, mean ± SEM for standard, and 1.19e-4 ± 4.59e-5 for ξ-input; $p<0.05$ for both pairwise comparisons with APE-trained according to an independent t-test, see **Fig. S2**). Thus, by predicting when errors might occur, neural systems learn to encode environment controllability in a useful format.

Together, these analyses imply that deep RL agents are able to learn an optimal meta-control policy, but only when asked to explicitly predict how controllable an environment is. However, this claim is potentially limited by our reliance on the frequency of observe choices as a proxy for the agents' estimate of controllability. Optimal behaviour on the task requires "observe" choices to be made more frequently in controllable environments, and so (for example) agents could be learning to make auxiliary (non-bet) actions whenever the world is more controllable. To rule this out, it becomes necessary to modify the task to include a condition where auxiliary actions were required when controllability was low, rather than high. This is what we did next.

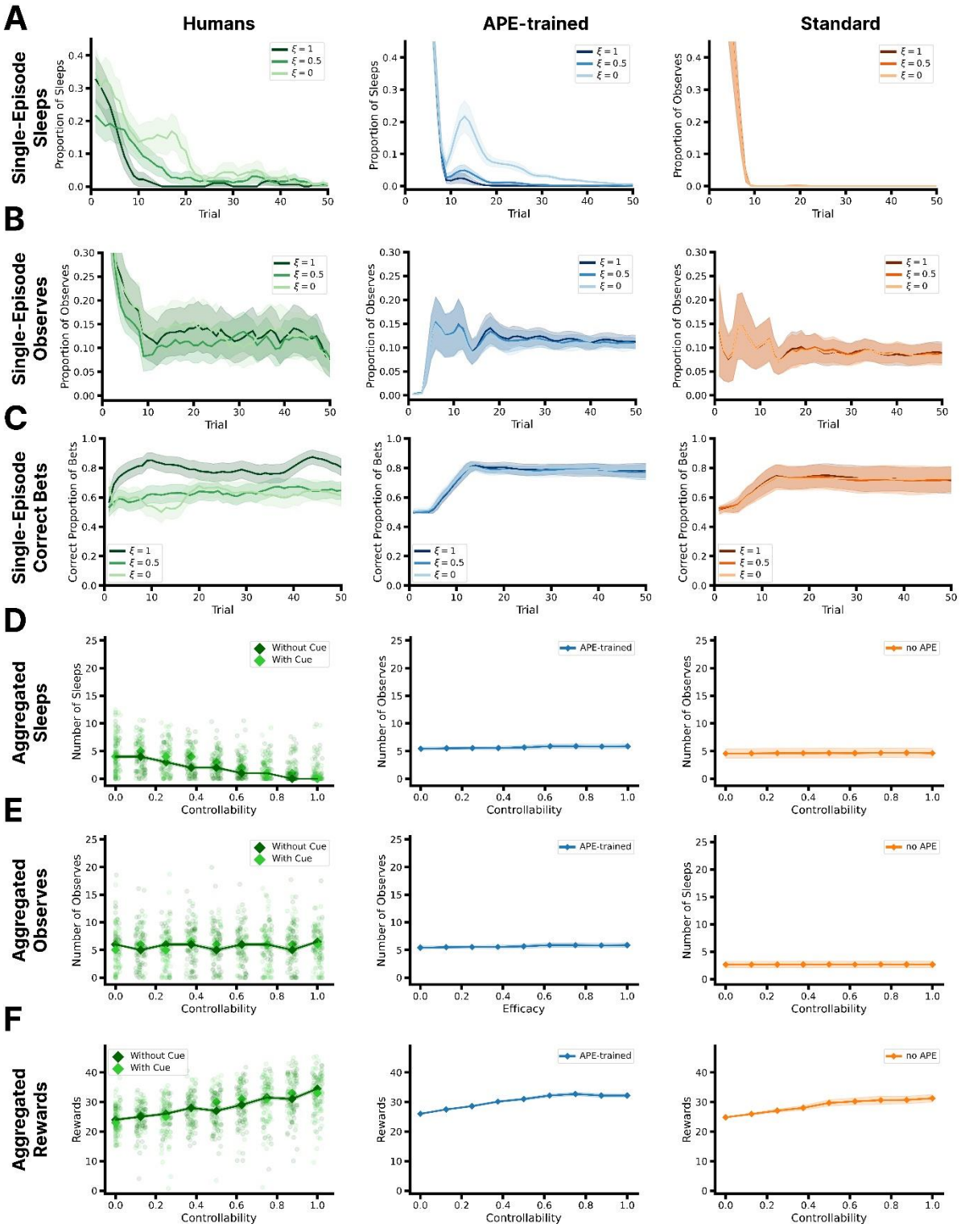

Figure 4. Humans can correctly solve the second *observe of bet* task (in a way that is well-modelled by the APE-trained networks). (**A**) Average frequency of sleeps over the course of each episode across (*left*) humans, (*middle*) APE-trained networks, and (*right*) no-APE networks for three different efficacy levels, (*dark*) $\xi = 1$, (*medium*) $\xi = 0.5$, (*light*) $\xi = 0$. Moving-average smoothed with a window of length 4. For the networks, the average is taken over the policy probabilities. Shaded region represents standard error of the mean. (**B**) Same plot at in (A), but for the frequency of observes. Moving-average smoothed with a window length of 8. (**C**) Same plot as in (B), probability of intending to place a bet on the right arm. (**D**) Average number of sleeps

across different efficacy levels for (*left*) humans, (*middle*) APE-trained networks, and (*right*) no-APE networks. For humans, individual samples are shown; for the networks, the shaded regions represent standard error of the mean. (**E**) Same plot as in (D), but for the average number of observations. (**F**) Same plot as in (C), but for the number of rewards obtained in an episode.

Learning to increase control via self-care

On day 3, human participants returned for an extra session in which they performed a new test (task 2). In this task they were offered (on each trial, in addition to observe and bet options) a third alternative that increased ξ by 0.1 for the remainder of the episode (to a maximum of 1). We call this the "sleep" option, and equate it with self-care processes that increase controllability and thus reduce response errors. For example, a tennis player will serve more accurately after a good night's rest. In this new task, participants learned to use the sleep action early on in training, and between trial 8-20 they used it more frequently when ξ was low (**Fig. 4A**, left panel), leading to an overall downward trend in use of the sleep button and the controllability of the environment (**Fig. 4D**, left panel) (average sleep rate: 0.005 ± 0.004, mean ± SEM, n=46 for ξ=1; 0.050 ± 0.011, n=65 for ξ=0.5; 0.134 ± 0.020, n=46 for ξ=0; $p<0.05$ for all pairwise comparisons according to Tukey's HSD; see Supplement for full statistics). This meant that just on day 2, overall human reward grew with ξ (total rewards: 33.83 ± 0.85, mean ± SEM, n=46 for ξ=1; 26.71 ± 0.69, n=65 for ξ=0.5; 24.02 ± 0.58, n=46 for ξ=0; $p<0.05$ for all pairwise comparisons according to Tukey's HSD; see Supplement for full statistics). By contrast, now the number of observe responses now did not grow with level of ξ (average observation rate starting from trial 15: 0.118 ± 0.010, mean ± SEM, n=46 for ξ=1; 0.107 ± 0.010, n=65 for ξ=0.5; 0.115 ± 0.11, n=46 for ξ=0; $p>0.82$ for all pairwise comparisons according to Tukey's HSD; see Supplement for full statistics) (**Fig. 4B** and **Fig. 4E**, left panels). This presumably occurs because the APE-trained networks learn to sleep consistently at the start of the block, effectively rendering every condition highly controllable, and obviating the need to observe differentially with levels of ξ.

As can be seen in **Fig. 4A-F** (centre panels), these findings were almost precisely recapitulated by the *APE-trained* network, including the time course of sleep choices for each level of environmental controllability (**Fig. 4A**). By contrast, the *standard* network did not learn to sleep more in less controllable environments (thus rendering them controllable; **Fig. 4A-F**, right panels). The APE-trained model's policy therefore predicted significantly better human choices than the standard model across efficacy levels (average likelihood: 0.661 ± 0. 030, mean ± SEM for APE-trained and 0.594 ± 0.024, mean ± SEM for standard, n=46, for ξ=1; 0.554 ± 0.031, mean ± SEM for APE-trained and 0.510 ± 0.025, mean ± SEM, n=65 for ξ=0.5; 0.594 ± 0.028, mean ± SEM for APE-trained and 0.535 ± 0.028, mean ± SEM for standard, n=46 for ξ=0; estimated model frequencies of 0.9955 ± 0.0006 and 0.0045 ± 0.0006, mean ± SEM, n=111 and exceedance probabilities of 1.000 and 2.161e-37 for APE-trained and standard, respectively; see Supplement for full statistics and Methods for details).

A deep network model of learned helplessness and compulsivity

Many psychological disorders can be conceived of as pathologies of control[4]. For example, people that underestimate the level of environmental controllability may refrain from taking useful actions in the belief that they are pointless. This is called learned helplessness, and it has been proposed to be a major component of depression and anxiety disorders[46,47]. By contrast, agents that overestimate the level of environmental controllability may compulsively over-explore, even when doing so has no effect on outcomes. This is reminiscent of the behavioural pathologies in obsessive-compulsive disorder (OCD), where participants may engage in stereotyped or ritualistic actions that lack apparent motive. OCD has been linked to misperception of control[48].

We used our APE-trained network to create simulations of these pathologies. At test, we added a perturbation to the hidden unit representation along the axis that corresponded to the trained controllability-readout of the APE-trained network, so that its output was systematically biased to over- or under-estimated environmental controllability. This neural perturbation was successful in increasing or decreasing the number of "observe" responses (total number of observes across all controllability settings: 53.59 ± 2.99, mean ± SEM, n=10 for positive ξ-perturbation; 48.45 ± 2.62, n=10 for no ξ-perturbation; 41.07 ± 2.66, n=10 for negative ξ-perturbation; $p<0.05$ for all pairwise comparisons according to a paired t-test with Holm-Bonferroni correction for multiple comparisons). Since the perturbation was introduced at inference time (test), the network continued to respond as if the recurrent layer was representing controllability truthfully. The results are shown in **Fig. 5A**, for both task 1 and task 2 (similar results were obtained by giving false feedback to the networks, explicitly "gaslighting" them into believing that the world was more or less controllable than it really was, see **Fig. S3**). In task 1, both positive and negative perturbation rendered the network relatively insensitive to environmental controllability, leading to a much shallower relationship between ξ and the frequency of 'observe' choices. However, when agents were (negatively) perturbed to underestimate controllability, they consistently failed to 'observe' and just harvested whatever rewards were possible by chance, reducing overall reward (a form of learned helplessness). By contrast, agents (positively) perturbed to overestimate controllability were excessively prone to observe, even when it was not reward-maximising to do so (according to the SARSOP baseline). This resembles compulsive behaviour, where patients repeatedly attempt to verify unobservable states (e.g. that front door is locked or gas is turned off), even when there is no apparent benefit to doing so.

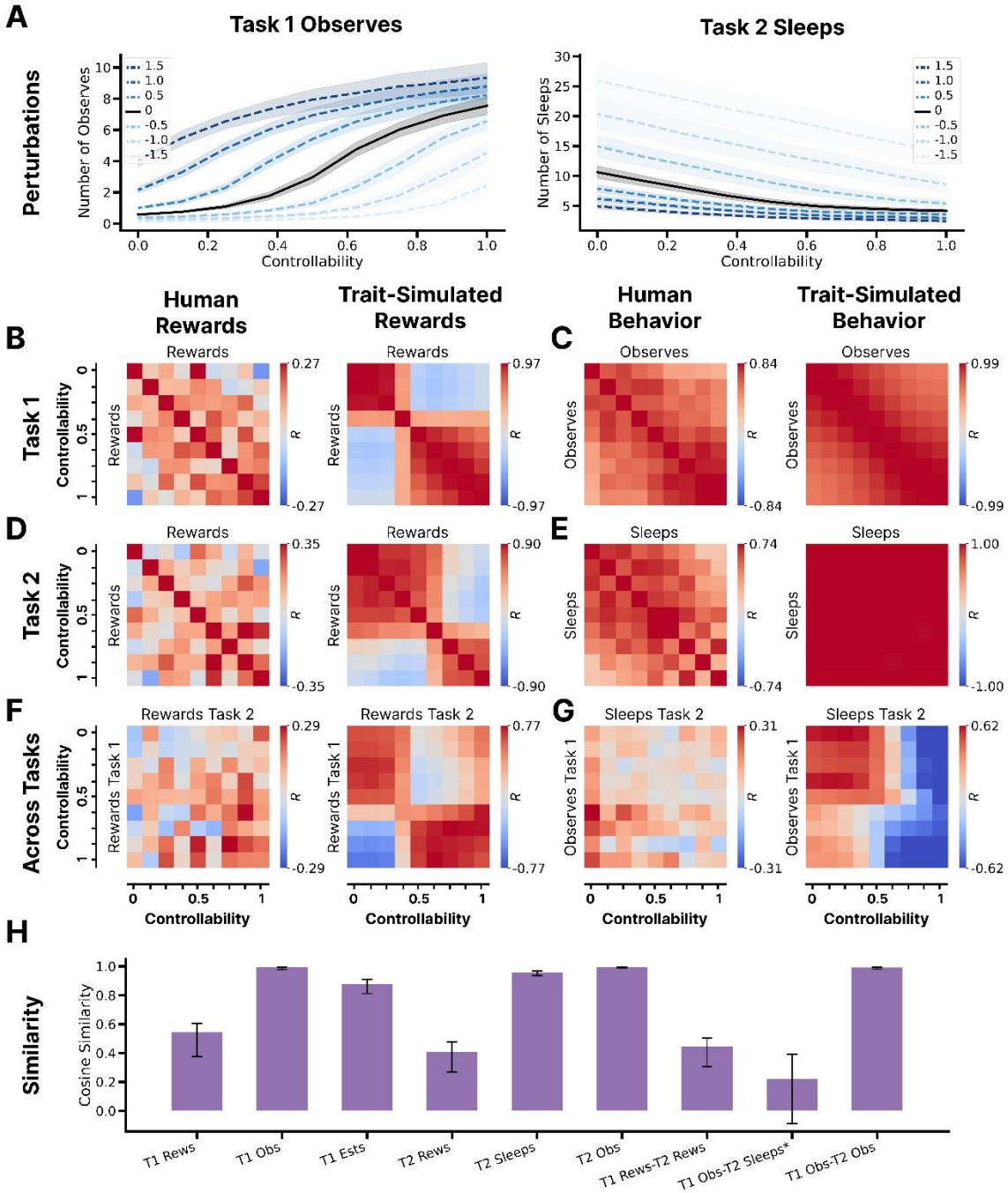

Figure 5. Neural networks can model individual differences in human performance. (**A**) The effect on the network policy when the population representation is perturbed to correspond to greater or lower values of efficacy by inverting the linear readout of the APE-trained network and pushing the population representation according to the corresponding projection, effectively causing the network to believe it has more or less control than it does in an episode. The intervention is performed on (*left*) task 1, with the average number of observes across tested controllability levels is showed for networks perturbed to (*lighter*) low and (*darker*) high levels of controllability, and (*right*) for task 2, where the average number of sleeps is shown. (**B**) The rewards received by participants in task 1 follows a stereotypical correlation structure in which participants' number of rewards is correlated across nearby controllability levels, but anti-correlated across controllability levels that are further away. The matrix shows the correlation between rewards obtained in task 1 by individuals across different levels of efficacy for (left)

humans and (*right*) "trait-level" simulated networks, i.e. those whose efficacy representation is perturbed by the same amount across different episodes. (**C**) Same as (B), but showing the correlation in number of observes across different episodes in task 1. (**D**) Same as (C), but showing the correlation in number of rewards obtained across different episodes in task 2. (**E**) Same as (D), but showing the correlation in number of sleeps across different episodes in task 1. (**F**) Same as (E), except showing the correlation from (*rows*) task 1 to (*columns*) task 2 of rewards obtained. Since these matrices are no longer symmetric, the whole matrix is used to compute the cosine similarity. (**G**) Same as (F), but showing the correlation between (*rows*) number of observes on task 1 and (*columns*) number of sleeps on task 2. Due to the correlation in observe actions across tasks, the number of observe actions on task 2 is partialled out of both variables. (**H**) Cosine similarity describing the degree of similarity between correlation matrices is between humans and the trait-simulated neural network agents. Upper-triangular matrices excluding the long diagonal are used to compute the cosine similarity when the matrix is an autocorrelation matrix. Error bars show 95%-confidence interval as computed using bootstrapping over 100 iterations.

In task 2, as expected, we see the converse pattern – networks that underestimate the controllability of the environment (i.e. depressive / anxious phenotype) tend to engage in excessive levels of 'sleep' across the board, whereas those that overestimate controllability (compulsive phenotype) fail to the "sleep" action irrespective of the value of ξ (total number of sleeps across all controllability settings: 37.32 ± 2.99, mean ± SEM, n=10 for positive ξ-perturbation corresponding to an increased belief in controllability, 58.27 ± 4.70, n=10 for no ξ-perturbation corresponding to an accurate belief in controllability; 127.90 ± 17.30, n=10 for negative ξ-perturbation corresponding to a lowered belief in controllability; $p<0.05$ for all pairwise comparisons according to a paired t-test with Holm-Bonferroni multiple comparisons correction; see Methods).

Next, we examined patterns of behaviour in a population of agents for which the level of perturbation was randomly sampled, creating a spectrum of individual differences that included both anxious/depressive (A/D) and compulsive network phenotypes. We analysed how consistent network behaviour was across different controllability levels for a given hidden layer perturbation. For each task as well as between tasks, we plotted the correlation in rewards (**Fig. 5B-F**, right panels) and observe or sleep probability (**Fig. 5C-G**, right panels) across the population, for each level of environmental controllability ξ with every other (in nine bins). In APE-trained networks, the cross-correlation plots showed a characteristic "blocked" structure for reward correlations, whereby there were positive correlations in reward for both low and high controllability blocks, but negative correlations between them. This was true for correlations within task 1 (**Fig. 5B**), within task 2 (**Fig. 5C**) and between tasks 1 and 2 (**Fig. 5D**, all right panels). This result implies that there are two pools of strategies in the neural network population: agents that yield higher rewards in less controllable environments fare more poorly in more controllable environments, and vice versa. Notably, these structured correlations in behavior were not shown by networks whose control representation was unperturbed, or was perturbed differently between different episodes (see **Fig. S4** and **Fig. S5** for these controls and the remaining behavioral variables).

When we plotted the same data for human participants, despite being more variable, the same approximate patterns emerged (**Fig. 5B-G**, left panels). To quantify their correspondence with networks, we computed the cosine similarity of the network cross-

correlograms with the human reward data, and found significant similarities (cosine similarity: 0.544, 90% CI = 0.375 to 0.604; **Fig. 5B**). Meanwhile, population similarity with networks that were perturbed with the same strength, but where the perturbations was shuffled between efficacy levels, was of significantly lower magnitude and inconsistent sign (see Supplement for statistics). We repeated this approach, correlating observe probability within task 1 (cosine similarity: 0.993, 90% CI = 0.977 to 0.996, original and 90% CI; **Fig. 5C**), sleep probability within task 2 (cosine similarity: 0.957, 90% CI = 0.936 to 0.968; **Fig. 5E**) and the correlation between rewards across tasks (cosine similarity: 0.444, 90% CI = 0.305 to 0.503; **Fig. 5G**) and observes in task 1 and sleeps in task 2 after controlling for the correlation with observes on task 2 (cosine similarity: 0.222, 90% CI: -0.089 to 0.392; **Fig. 5G**). In all cases, we saw greater similarity than would be expected by chance, although the difference was not significant for observes and sleeps (**Fig. 5H**). In other words, the overall patterns of behaviour and outcomes across the population varied as a function of environmental controllability in a mostly conserved fashion between humans and APE-trained networks.

Transdiagnostic scores predict how individuals learn the value of control

In the human population, we saw patterns of correlation that resembled those in APE-trained networks that were perturbed to either over- or under-estimate the level of controllability of the environment. This is presumably because our randomly sampled human population contains a spectrum of individual differences, with some participants more prone to depressive/anxious type phenotypes, and others to show forms of compulsivity. With this idea in mind, at the end of the study we asked participants to complete a short self-report questionnaire which has successfully used to capture these traits in healthy populations[33] (using scoring tools available at https://github.com/the-wise-lab/FACSIMILE/). The survey required participants to note how much they agreed with statements such as “I get tired for no reason” and “I get upset if things are not arranged properly”, and yielded scores on three transdiagnostic dimensions: (1) anxiety and depression, (2) compulsivity and intrusive thoughts, and (3) social withdrawal[49]. We pre-registered our prediction that participants with higher scores on Factor 1 would be less prone to observe in task 1, and more prone to sleep in task 2, whereas the converse would be true for those with higher scores on Factor 2 (we had no predictions either way about social withdrawal).

In **Fig. 6A**, we show the predictions of a linear regression in which A/D and compulsivity scores were included as joint predictors of observation probability in task 1 (left) and sleep probability in task 2 (right). The reward-maximising strategy is highly dependent on ξ, so we perform the analysis separately for both low controllability levels with $\xi \in \{0, 0.125, 0.25, 0.375\}$ and high controllability levels with $\xi \in \{0.625, 0.75, 0.875, 1\}$. Contrary to our pre-registered predictions, we did not see convincing evidence for a linear relationship between transdiagnostic scores and the probability of observe choices. In task 1, the resulting coefficient for A/D was weakly positive but not significant for either high or low controllability levels (coefficient m=0.592, p=0.121, n=111 for low ξ; m=0.623, p=0.100, n=111 for high ξ). The relationship for compulsivity was weakly negative for low controllability settings, and weakly positive for high controllability (m=0.-0.339, p=0.384, n=111 for low ξ; m=0.058, p=0.881, n=111 for high ξ). In task 2, the linear

relationships were weakly and non-significantly positive for all settings (AD: m=0.210, p=0.363 for low ξ; m=0.194, p=0.345 for high ξ; compulsivity: m=0.089, p=0.704 for low ξ; m=0.143, p=0.495 for high ξ).

We observed that A/D and compulsivity scores were positively correlated (r^2=0.358, p=1.16e-4; **Fig. 6B**), which could be indication of a shared axis of psychopathology. Since our neural network simulations predict opposite effects of A/D and compulsivity phenotypes on estimated controllability, we adopted an exploratory approach in which we identified factors of variance in questionnaire data using data-driven methods. We orthogonalized the space spanned by A/D and compulsivity by calculating the principal components across both factors. The first principal component (PC1), which had positive coefficients across both factors (0.729 for AD and 0.685 for compulsivity) and explains 67.9% variance, presumably represents generic psychopathological factors shared across the whole cohort. However, the second principal component (PC2), which we orient to have a positive coefficient for A/D (0.685) and a negative coefficient for compulsivity (-0.729), was a better candidate to index the transdiagnostic symptoms that differ between these phenotypes. In **Fig. 6E**, we plot the joint distribution of observe choices as a function of controllability and quantile of PC1 (left panels) and PC2 (centre left panels) on task 1, and in **Fig. 6H** for task 2.

In the centre right panel, we plot the predictions from the trait-simulated deep network population, sorted by quantile of perturbation. Each plot shows how a given variable (e.g. rewards in **Fig. 6D**) varies jointly with both controllability (on the x-axis) for participants sorted into noniles by their transdiagnostic scores, as measured by PC2 (see **Fig. S6** for similar effects on PC1, **Fig. S7** for original transdiagnostic scores, and Supplement for full statistics). On the rightmost panel we see the predictions from an unperturbed network, which by definition has the same response for each quantile – as there is no variation by trait. The trait-simulated network predicts that several variables will depend not just on controllability but on transdiagnostic score, or on the interaction between these variables. As can be seen, transdiagnostic scores interact strongly with behavioural performance across controllability levels: Participants with higher PC2 score tend to observe slightly more frequently on task 1 (higher values in the lower part of the plot) and strongly more on task 2, matching the population response of the perturbed APE-trained networks (fraction remaining similarity explained by the trait-simulated networks: 0.015, 90%-CI = -0.183 to 0.154 for task 1; 0.557, 90%-CI = 0.618).

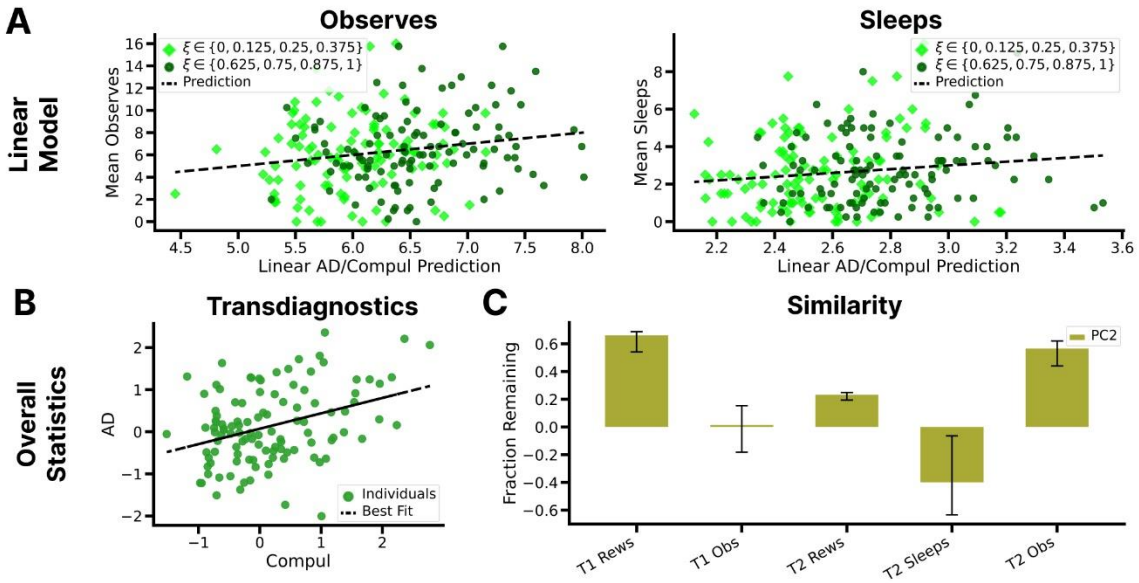

Figure 6 (Part 1). Neural networks can predict behavioural effects of transdiagnostic factors. (**A**) True versus predicted human behaviour for a linear model trained to predict (*left*) mean number of observes per individuals on task 1 and (*right*) mean number of sleeps in task 2, based on their A/D and compulsivity transdiagnostic factors, computed using the questionnaire and methods from Hopkins et al (2022). The linear models are fitted separately on low-controllability episodes (*light green*, $\xi \in \{0, 0.125, 0.25, 0.375\}$) and high-controllability episodes (*dark green*, $\xi \in \{0.625, 0.75, 0.875, 1\}$. The linear model is statistically significant ($p<0.05$) based on an f-test for both linear models trained to predict number of observes and insignificant for the model trained to predict number of sleeps. The points correspond to values for individual people, and the dotted black line shows the identity where the true value matches the prediction exactly. (**B**) Relationship between A/D and compulsivity for (*individual participants*) and (*dashed* grey) the line of best fit. (**C**) Difference in cosine similarity for the matrices between the trait-level perturbations, 0-perturbation networks, and human behaviour based on transdiagnostic factors A/D and compulsivity. Error bars show 95%-confidence interval as computed using bootstrapping over 100 iterations.

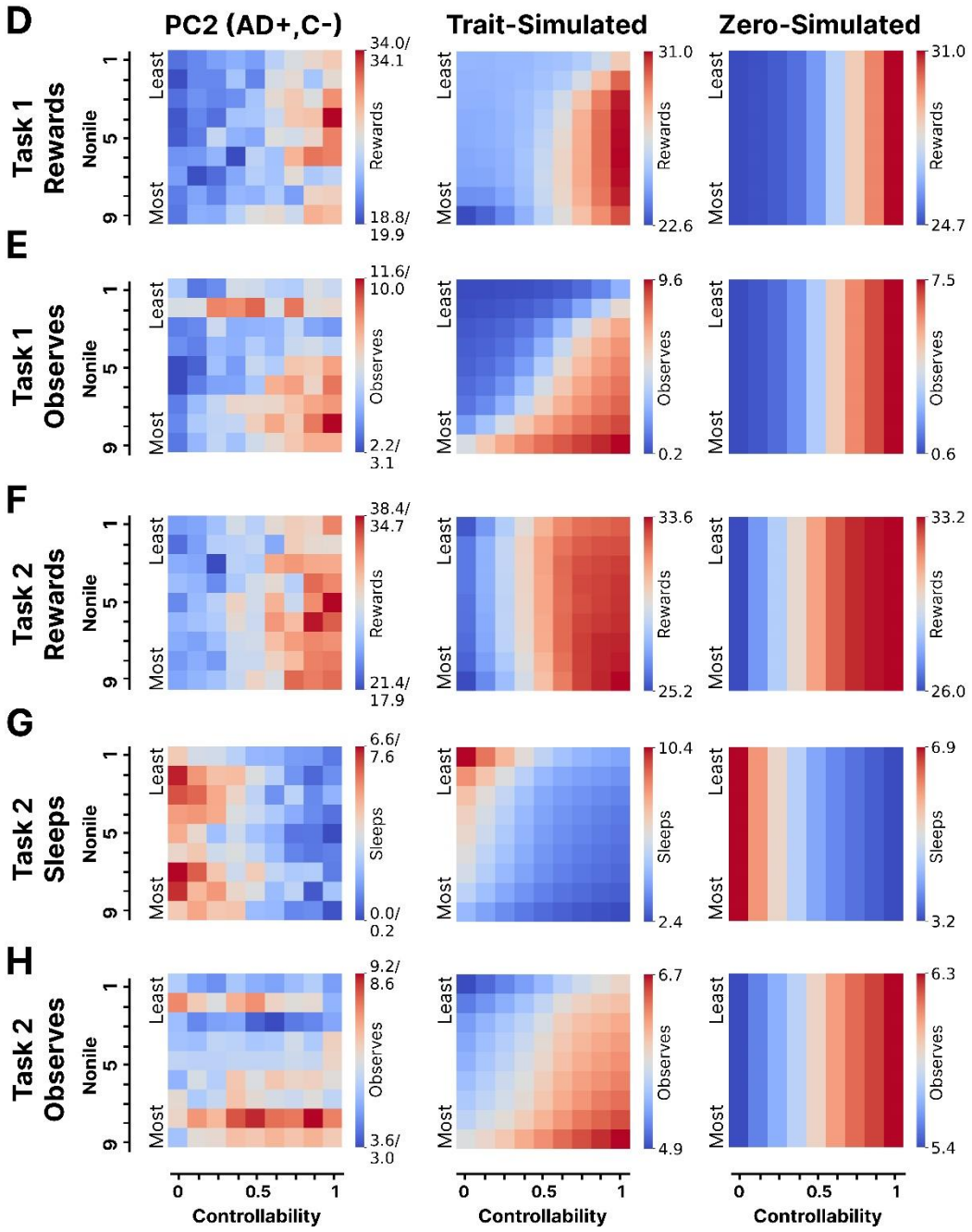

Figure 7 (Part 2). Neural networks can predict behavioural effects of transdiagnostic factors. (**D**) Mean reward tallies in task 1 across individuals within (*row*) a particular nonile for episodes in (*column*) a given efficacy level, where the noniles are taken according to transdiagnostic factor scores (*far left*) AD, (*middle left*) compulsivity, and (*middle right*) perturbation magnitude for neural networks, with (*far right*) representing a control showing the values for an unperturbed neural network. The matrices showing human performance are computed separately and normalized between minimum and maximum values for each of the two groups, and then combined using a weighted average based on the corresponding number of participants. (**E**) Same as (D), but showing the number of observe actions taken per episode in task 1. (**F**) Same as (E), but showing the reward tallies per episode in task 2. (**G**) Same as (F), but showing the number of sleep actions obtained per episode in task 2. (**H**) Same as (G), but showing the number of sleep actions obtained per episode in task 2.

Over the other panels in B-F we repeated this exercise for relevant variables, namely rewards (**Fig. 6B**) in task 1 and rewards (**Fig. 6E**) and sleep (**Fig. 6F**) actions in task 2 (see **Fig. S5** for the remaining variables). PC2 was correlated significantly more strongly with behavior of the perturbed neural networks for all of these factors except number of sleep actions in task 2 (fraction remaining similarity explained: 0.662, 90%-CI = 0.542 to 0.687 for task 1 rewards; 0.232, 90%-CI = 0.194 to 0.249 for task 2 rewards; -0.399, 90%-CI = -0.632 to -0.063 for task 2 sleeps, see **Fig. 6C**). The relationship between transdiagnostic scores and performance in the task is further highlighted by the difference in transdiagnostic scores between participants included in the study and those who were excluded from analysis on the basis of pre-registered criteria for out-of-distribution behaviour on the practice day (A/D: 0.135 ± 0.084, mean ± SEM, N=111 for included and -0.199 ± 0.107, mean ± SEM, N=71 for excluded, $t(180)=2.437$, $p<0.05$; compulsivity: 0.171s ± 0.082, mean ± SEM, N=111 for included and -0.043 ± 0.110, mean ± SEM, N=71 for excluded, $t(180)=1.546$, $p=0.124$). Together, these results imply that our perturbed neural networks were able to capture human variation in a tendency to under- or overestimate the controllability of the environment, as measured by responses to transdiagnostic questionnaires. However, the relationship between survey responses and the predicted patterns of behaviour is our task is more complex than the one that we originally pre-registered, and will need to be confirmed with further research.

Discussion

We studied the ability of humans and deep neural networks to learn a meta-control policy – to adapt their behaviour to the level of controllability of the environment, operationalised as the probability that an action is converted into its intended consequences. We did not explicitly cue environmental controllability in either humans or networks, who instead had to estimate this quantity on the fly from the history of actions and outcomes. For computational inspiration, we drew upon the framework of deep meta-learning, which is increasingly popular as a theory of human cognition[38,39]. Deep networks were able to meta-learn a near-optimal solution to this task, but only when explicitly trained to the likelihood of an error – and thus to infer levels of environment controllability.

This finding provides a strong computational explanation for why biological agents have evolved to explicitly monitor and update levels of uncertainty or conflict in the environment, or the likelihood than an error will occur. It also explains why APEs – a neural signal coding mismatch between intended and realised actions – are such prominent brain responses for humans and other animals[23–26,28,32]. One of the most interesting aspects of our data was that it was not sufficient for the network simply to represent environmental controllability *per se*. In fact, in ξ-input networks, which did not compute APEs but instead received controllability explicitly as an input, this variable was consequently much easier to decode – and yet these networks never learned an optimal meta-control policy. In other words, there exists a deep connection between the requirement to predict how controllable the environment is, and the capacity to reformat information in a way that permits the exercise of cognitive control.

Our human behavioural experiments show that the APE-trained network learned a meta-control policy that was very similar to that displayed by people. Figures 2 and 4 show strikingly conserved behavioural signatures (between humans and networks) over the course of blocks with different levels of controllability. Of note, our manipulation involved both conditions where lower levels of environmental controllability required participants to overlook ('observe' in task 1) or engage with ('sleep' in task 2) auxiliary actions, and so are unlikely to be driven by generalised levels of engagement or disengagement with the task. Rather, both humans and APE-trained networks learn how to exploit available controllability (by observing) or to increase controllability (by sleeping) in the service of reward-maximisation.

We used perturbations to create deep networks that exhibit individual differences in how they learn the value of control – being encouraged to either over- or under-estimate environmental controllability. We found that human participants who report transdiagnostic traits previously associated with anxiety / depression and compulsivity display patterns of behaviour that are similar to neural networks that are perturbed to over- and under-estimate environmental controllability respectively (as derived from the sign of the coefficients on the second principal component). This is consistent with theories that cast these pathologies as disorders of control that result in behavioural abnormalities. For example, depressed patients often neglect to take the initiative – to seek information that would permit reward-maximising behaviours, which could come from under-estimating the likelihood that these actions will translate to a positive outcome as intended. Conversely, OCD patients may over-estimate the controllability of the environment, leading them to compulsively check for the value of latent states, at the expense of reward-maximising. Both perturbations thus create behavioural pathologies.

However, we note here that the correspondence between over- and under-estimation of environmental controllability and the symptomatology of A/D and compulsion is not precise. In particular, we found that the main component of transdiagnostic scores in our convenience sample was a generic one which jointly indexed both A/D and compulsive symptomatology. The axis spanning an A/D phenotype (indicative of underestimation of environmental controllability) and a compulsive phenotype (over-estimation) emerged in the second principal component, and its relationship with the predictions of our population of perturbed neural networks was close but not exact. It is possible that with a significantly larger cohort (thousands rather than ~100 participants) this result would be much cleaner. Nevertheless, we think that our work opens up new avenues for studying these behaviours using computational tools based on deep neural networks.

References

1. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* **24**, 167–202 (2001).
2. Shiffrin, R. M. & Schneider, W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review* **84**, 127–190 (1977).
3. Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S. & Cohen, J. D. Conflict monitoring and cognitive control. *Psychol Rev* **108**, 624–52 (2001).
4. Huys, Q. J. M. & Dayan, P. A Bayesian formulation of behavioral control. *Cognition* **113**, 314–328 (2009).
5. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–40 (2013).
6. Lee, J. H., Leibo, J. Z., An, S. J. & Lee, S. W. Importance of prefrontal meta control in human-like reinforcement learning. *Front. Comput. Neurosci.* **16**, 1060101 (2022).
7. Eppinger, B., Goschke, T. & Musslick, S. Meta-control: From psychology to computational neuroscience. *Cogn Affect Behav Neurosci* **21**, 447–452 (2021).
8. Richards, B. A. *et al.* A deep learning framework for neuroscience. *Nat Neurosci* **22**, 1761–1770 (2019).
9. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat Rev Neurosci* **22**, 55–67 (2021).
10. Doerig, A. *et al.* The neuroconnectionist research programme. *Nat Rev Neurosci* vol. 24 431–450 (2023).
11. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci* **1**, 417–446 (2015).

12. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–44 (2015).
13. Yamins, D. L. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–65 (2016).
14. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
15. Orhan, A. E. & Ma, W. J. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nat Neurosci* **22**, 275–283 (2019).
16. Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J. & Wang, X. J. Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions. *Neuron* **93**, 1504-1517 e4 (2017).
17. Mnih, V. *et al*. Human-level control through deep reinforcement learning. *Nature* **518**, 529–33 (2015).
18. Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J. & Kurth-Nelson, Z. Deep Reinforcement Learning and Its Neuroscientific Implications. *Neuron* **107**, 603–616 (2020).
19. Sandbrink, K. & Summerfield, C. Learning the value of control with Deep RL. in *2023 Conference on Cognitive Computational Neuroscience* (Cognitive Computational Neuroscience, Oxford, UK, 2023). doi:10.32470/CCN.2023.1640-0.
20. Bustamante, L., Lieder, F., Musslick, S., Shenhav, A. & Cohen, J. Learning to Overexert Cognitive Control in a Stroop Task. *Cogn Affect Behav Neurosci* **21**, 453–471 (2021).
21. Siqi-Liu, A. & Egner, T. Contextual Adaptation of Cognitive Flexibility is driven by Task- and Item-Level Learning. *Cogn Affect Behav Neurosci* **20**, 757–782 (2020).

22. Dey, A. & Bugg, J. M. The Timescale of Control: A Meta-Control Property that Generalizes across Tasks but Varies between Types of Control. *Cogn Affect Behav Neurosci* **21**, 472–489 (2021).
23. Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E. & Donchin, E. A Neural System for Error Detection and Compensation. *Psychol Sci* **4**, 385–390 (1993).
24. Falkenstein, M., Hohnsbein, J., Hoormann, J. & Blanke, L. Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology* **78**, 447–455 (1991).
25. Carter, C. S. *et al*. Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance. *Science* **280**, 747–749 (1998).
26. Holroyd, C. B. & Coles, M. G. H. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review* **109**, 679–709 (2002).
27. Narayanan, N. S., Cavanagh, J. F., Frank, M. J. & Laubach, M. Common medial frontal mechanisms of adaptive control in humans and rodents. *Nat Neurosci* **16**, 1888–1895 (2013).
28. Greenstreet, F. *et al*. *Action Prediction Error: A Value-Free Dopaminergic Teaching Signal That Drives Stable Learning*. http://biorxiv.org/lookup/doi/10.1101/2022.09.12.507572 (2022) doi:10.1101/2022.09.12.507572.
29. Bogacz, R. Dopamine role in learning and action inference. *eLife* **9**, e53262 (2020).
30. Gratton, G., Coles, M. G. & Donchin, E. Optimizing the use of information: strategic control of activation of responses. *J Exp Psychol Gen* **121**, 480–506 (1992).

31. Rabbitt, P. M. Errors and error correction in choice-response tasks. *Journal of Experimental Psychology* **71**, 264–272 (1966).
32. Brown, J. W. & Braver, T. S. Learned predictions of error likelihood in the anterior cingulate cortex. *Science* **307**, 1118–21 (2005).
33. Wise, T., Robinson, O. J. & Gillan, C. M. Identifying Transdiagnostic Mechanisms in Mental Health Using Computational Factor Modeling. *Biological Psychiatry* **93**, 690–703 (2023).
34. Tversky, A. & Edwards, W. Information versus reward in binary choices. *Journal of Experimental Psychology* **71**, 680–683 (1966).
35. Navarro, D. J., Newell, B. R. & Schulze, C. Learning and choosing in an uncertain world: An investigation of the explore–exploit dilemma in static and dynamic environments. *Cognitive Psychology* **85**, 43–77 (2016).
36. Blanchard, T. C. & Gershman, S. J. Pure correlates of exploration and exploitation in the human brain. *Cogn Affect Behav Neurosci* **18**, 117–126 (2018).
37. Kurniawati, H., Hsu, D. & Lee, W. S. SARSOP: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces. in *Robotics:Science and Systems IV* (MIT Press, 2009).
38. Wang, J. X. Meta-learning in natural and artificial intelligence. *arXiv:2011.13464 [cs]* (2020).
39. Binz, M. *et al.* Meta-Learned Models of Cognition. Preprint at http://arxiv.org/abs/2304.06729 (2023).
40. Williams, R. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *MachLearn* **8**, 229–256 (1992).

41. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735–1780 (1997).

42. Hazy, T. E., Frank, M. J. & O'Reilly, R. C. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Phil. Trans. R. Soc. B* **362**, 1601–1613 (2007).

43. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* **21**, 860–868 (2018).

44. Duan, Y. *et al.* RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv* (2016).

45. Shenhav, A., Botvinick, M. M. & Cohen, J. D. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–40 (2013).

46. Abramson, L. Y., Metalsky, G. I. & Alloy, L. B. Hopelessness depression: A theory-based subtype of depression. *Psychological Review* **96**, 358–372 (1989).

47. Maier, S. F. & Seligman, M. E. P. Learned helplessness at fifty: Insights from neuroscience. *Psychological Review* **123**, 349–367 (2016).

48. Reuven-Magril, O., Dar, R. & Liberman, N. Illusion of control and behavioral control attempts in obsessive-compulsive disorder. *Journal of Abnormal Psychology* **117**, 334–341 (2008).

49. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).

Methods

Participants

Data collection was carried out online using the platform Prolific Academic (https://prolific.com). The experimental was collected in two batches over successive weeks using a three-day procedure, with 110 participants recruited each week. We only analyzed data from those participants who completed the full 3-day experiment, including the transdiagnostic psychiatric questionnaire at the end. In the end, we had full data from 182 participants across both weeks and groups. Of these participants, some almost never used the "observe" option (2 occasions or fewer), and some used it almost exclusively (18 occasions or more). We excluded these according to a pre-registered procedure, leaving 111 participants.

When they joined the experiment, participants were randomly assigned to either group A or group B. Of the 182 participants, 86 were in group A and 96 in group B. Of the 111 participants following exclusions, 46 were in group A and 65 were in group B.

Participants were compensated for both their time and for correct choices. As base pay, participants received 5 GBP for day 1 (estimated completion time: 40 minutes), 3 GBP for day 2 (estimated completion time: 25 minutes), and 3.75 GBP for day 3 (estimated completion time: 30 minutes) for their time. Additionally, participants received an 0.01 GBP for each point they earned during the game as a bonus. Average reward rate was around 9 GBP per hour on all days and for both batches.

All participants provided informed consent before taking part, and the studies were approved by the University of Oxford Central University Research Ethics Committee.

Stimuli, task and Procedure

On each trial of task 1, participants were shown a pair of blue and red circles symbolizing the betting actions, an icon of a pair of glasses for the observe action. The blue circle was on top, in the middle the glasses, and on the bottom the red circle. All had reduced opacity. In lieu of a fixation cross, the stimuli were first presented for 350 ms without additional information and without the option for participants to click on anything. After that, additional text appeared on top of the screen indicating the trial number as well as the phrases "Bet: Click on the light of your choice" and "Observe: Click on the glasses". At the bottom of the screen, the instructions "Click on one of the pictures to continue" appeared. Collectively, these indicated to participants that the intertrial interval was over and they could make an action. Participants chose one of the two betting actions (circles) or the glasses (observe) by directly clicking on the corresponding part of the stimulus. Participants could take as much time as they needed to click on an action. Stimuli remained on the screen until an action was selected. On trials in which participants selected a bet, the selected circle was modified by a dark outer ring to indicate that the light had been chosen by the participant, and an icon of a coin with a red question mark appeared to indicate where the participants' bet had been placed. These remained for 1000 ms. On observe trials, the glasses increased in opacity to indicate that an observe trial had been taken for 1000 ms, as did the circle corresponding to the light that would have paid out that round. At that point, the next trial began, again

with the same 350 ms intertrial interval during which the stimuli were shown but could not be selected. Stimuli were generated using a vector graphics editor. and arranged using JSPsych (de Leeuw et al., 2023). Additionally, participants saw a background image for the story intro featuring a casino table, generated by an AI (DALL-E).

At the end of an episode, participants were told how much reward they had earned. On training episodes, they additionally saw complete information about the entire trajectory of the episode, including which light lit up on which trial, which actions they intended to choose, and which actions were selected. On days 1 and 2, participants were additionally asked to indicate the level of control they had had during the episode using a slider.

Task 2 was identical, except that the action stimuli featured an icon of a bed on top to represent the sleep action, and below the blue circle, glasses, and red circle arranged horizontally from left to right. When selected, the bed action increased in opacity.

All three days were run in forced full screen mode. Day 1 was a practice day for task 1. It began with a short introductory story about the participant entering a casino and being explained the rules by a staff member at a table. The story intro featured the casino background image. This story was followed by plain-text explanations of the rules, that featured examples of moves that participants had to execute, and visualizations of the feedback that they received. These were on a white background. The instructions included quizzes on the instructions that were shown once, regardless of the correctness of participants' responses, but which were followed by feedback explaining why a participants' choice was right or wrong. To aid in participant comprehension, the instructions also explicitly described the most rewarding strategy as being to observe more in episodes where the participant had more control.

After that, participants played nine blocks of the game, one with each controllability level in the set $\{0.125\,k \mid 0 \leq k \leq 8, k \in \mathbb{Z}\}$. For group A, controllability levels $\{0, 0.25, 0.75, 1\}$ was used as the training set on day 1 and $\{0.125, 0.375, 0.5, 0.625, 0.875\}$ as test. For group B, the sets were exactly inverted. Within a set, the order of the controllability levels was randomized for each participant. Both training and test blocks used the same stimuli described above. However, on training sets, before the first trial, participants received a verbal description of the level of control they were going to have, "none" for ξ=0, "a little" for $0 < \xi < \frac{1}{3}$, "some" for $\frac{1}{3} \leq \xi < \frac{2}{3}$, "a lot of" for $\xi \geq \frac{2}{3}$. This was accompanied by a qualitative description of how often their bets would succeed: "half the time", "slightly more than half the time," "most," "almost always", and "always," respectively. Additionally, on training episodes, the screen's background color was changed to function as a controllability cue. The color was interpolated linearly between red for no control to yellow for complete control. The background color was changed at the moment participants received their verbal controllability description and remained that way for the remainder of the episode. In contrast, for the test episodes, participants received no such verbal description, and the background color was set to white.

On Day 2, the data was collected for task 1. It included a short recap of the instructions of the previous day on the casino background before switching back to a white

background for quizzes. Participants performed 9 blocks of 50 trials each. Again, each block involved a different controllability level. However, the allocation of controllability levels as training and test was now flipped. On this day, for group A, $\{0, 0.25, 0.75, 1\}$ was used as the training set on day 1 and $\{0.125, 0.375, 0.5, 0.625, 0.875\}$ as test. As on the previous day, the order was reversed for group B. This time, the training sets were not accompanied by a verbal cue and explanation of the controllability level. The background color was still changed to reflect controllability level during the training blocks as before. On test blocks, participants did not receive any indication of the level of control, and played on a white background.

Finally, on day 3, participants participated in task 2. The instructions began with a description of the added action again on the casino background image. They then reverted to plain-text instructions, The instructions also explicitly described the most rewarding strategy as being to sleep more in episodes where the participant had less control. Participants then again played nine blocks of 50 trials. As on day 2, for group A, $\{0, 0.25, 0.75, 1\}$ was used as the training set on day 1 and $\{0.125, 0.375, 0.5, 0.625, 0.875\}$ as test. These allocations were flipped for group B. On test blocks, participants did not receive a verbal description of the level of control. However, the background color was again interpolated between red and yellow as a cue for increasing controllability level. As the controllability level could change during a block in task 2, the background color was dynamically adjusted when participants chose the sleep action to always reflect the current level of control participants had. On test blocks, participants played on a white background and did not receive any indication of the level of controllability.

At the end of day 3, participants completed psychiatric transdiagnostic questionnaires from Hopkins et al. (2022), which allows for robust estimation of the three transdiagnostic psychiatric factors originally recorded in Gillan et al. (2016). The questions all consistent of statements that participants agreed with by selecting one of a number of options ranging from “Strongly Disagree” to “Strongly Agree.”

The experiment ended with a brief outro story again on the casino background.

The task and instructions are available online. The link for the first day is: https://www.hipexperiments.co.uk/sandbox/tasks/kai/HumanObserveBetEfficacyDay1Of2.

***Observe or bet* task settings**

Different versions of the task were used to compare human and neural network performance as well as to evaluate machine learning performance. The *observe or bet* tasks used in this report were all 50 steps long. In the second task, sleeping increased controllability by 0.1 for all task versions. No penalty was deducted from participants for placing a bet on the wrong bandit in order to allow a clearer disentangling of observation policies between different efficacy levels. The parameter settings that differed between versions are listed in Table 1.

Table 1. Parameters used in the *observe or bet* tasks.

Version	Task	Bias	Reversal probability
Human	Task 1	0.5	10%
Human	Task 2	0.5	5%
Machine Learning	Task 1	0.4	5%
Machine Learning	Task 2	0.49	5%

Calculation of optimal solutions

A partially-observable Markov decision process (POMDP) is a 7-tuple $\langle S, A, \Omega, O, T, r, \gamma \rangle$, where S is the (finite) non-empty state space, A is the (finite) non-empty action space, Ω is the (finite) non-empty observation space, $O : S \rightarrow P(\Omega)$ is the observation function, $T : S \times A \rightarrow P(S)$ is the probabilistic state-transition function, $r : S \times A \rightarrow P(R)$ is a bounded reward function, and $0 \leq \gamma \leq 1$ is the discount factor. We formalized the observe or bet task for the calculation of optimal solutions by defining the set of states as the product space given by the number of steps along with the two possibilities for which is the high-paying door. For task 1, the set of actions is to observe, to bet on the left door, or to bet on the right door. The set of observations are given by the product of the set of the number of steps and the possible observations per step, no observation, observing a payout on the left door, and observing a payout on the right door. The observation, transition, and reward functions that correspond to the regular observe-or-bet rules. We set $\gamma = 1$. This formulation of the POMDP takes the perspective of an agent who knows what the overall probability level is, but does not know the assignment to a particular door. We used this formulation to calculate the normative solutions, providing an upper bound on performance. In contrast, the neural networks are meta-trained across all probability levels, which means they start an episode with a uniform prior about the probability level. Because we only give participants general indication of probability ranges rather than an exact estimate, we can assume that they are and in any case start the trial without information about which arm is the higher-paying one, we can assume that the participants operate with some uncertainty as to the exact transition structure of the environment and therefore are operating under the POMDP described above.

We used the SARSOP algorithm to closely approximate the optimal solution to the decision-making task from the perspective of an agent that has knowledge of the transition structure of the environment, including the level of efficacy. SARSOP is a state-of-the-art solver for problems that require active information gathering (Ma and Pineau, 2015; Silver and Viness 2010). We used the implementation in the JuliaPOMDP package in the Julia programming language (Egorov et al. 2017). This solver generates the observations based on the state reached by the network after state transition, as opposed to the task given to the humans and the networks in which the observation was generated before the transition. This means that observe actions were more valuable for the POMDP solver than for the humans and networks, making the upper bound stricter. In practice, however, this difference is small.

Training and evaluation of neural network policies

We trained the APE-based and the solely-reward-based networks over a hierarchical task structure over ξ to maximize returns using meta-RL (Wang et al., 2016). All networks were

trained using REINFORCE with a baseline of 1/3 (corresponding to the expected value of a random action, as on every step exactly one bandit pays out and there are three possible actions). We used the policy gradient over the policy π to update network parameters θ. For a sampled Monte Carlo return $G_t = \sum_{t' \geq t}^{T} \gamma^{t'-t} \left(R_{t'} - \frac{1}{3} \right)$ and total number of trials in an episode T, the policy gradient for a trial *t* is:

$$\nabla_\theta \mathcal{L}_\pi = -G_t \cdot \nabla_\theta \ln \pi_\theta(\hat{a}_t | o_t)$$

The base neural network architecture contained an LSTM layer of 48 units, followed by a feedforward layer of 24 units, and finally a softmax output of the policy over the three possible actions. The LSTM activations were reset at the start of every episode. The error likelihood signal of the APE-trained networks consisted of a single node with a linear activation function that was connected fully to the LSTM layer. It was trained to predict the firing of APEs using a squared error loss function which trains the network to predict the error likelihood $\hat{p}_t$, resulting in the gradient for trial *t*:

$$\nabla_\theta \mathcal{L}_{\hat{p}} = \left[\hat{p}_t - \delta_{\mathrm{APE},t} \right] \cdot \nabla_\theta \hat{p}_t$$

with a weighting coefficient of strength 25 for task 1 and 100 for task 2. This node was then concatenated back to the LSTM layer so that it reached the next fully-connected layer as an additional input. The inputs to the network consisted of one-hot encodings of the previously-intended action $\hat{a}_{t-1}$ and the actually-taken action a_{t-1} inputs as well as a scalar indicating the fraction of steps that were remaining in the episode, a unit for each of the two lights that was set to 1 if an observe action was chosen and that light lit up on the previous turn, and, matching the human input, a binary flag and a scalar tally for rewards earned in the previous episode that were activated on the start of a new episode. The input to the ξ-input networks contained ξ as an additional scalar node.

We meta-trained the networks over the set $\xi \in \left[0, \frac{1}{3}\right] \cup \left[\frac{2}{3}, 1\right]$ and tested them on discrete values across the entire range $\xi \in \{0.125\, k \mid 0 \leq k \leq 8, k \in \mathbb{Z}\}$. We trained 10 instantiations per type over 500k episodes of 50 steps with a batch size of 50 episodes, annealing entropy regularization to 0 geometrically over 150k episodes. We used PyTorch without additional libraries (Paszke et al., 2019). No hyperparameter optimization was conducted beyond initial hand-tuning of parameters to reasonable values.

Analysis of neural networks

To determine how well efficacy is represented in the networks, we performed a population-level decoding analysis was represented of the recurrent layer's hidden units. In this analysis, the activations of all neurons in a hidden layer at a single time point were jointly used as predictors of efficacy at that point in time (constant over the episode). The accuracy was evaluated using r^2 score for time-points in the second half of an episode (to give the network time to determine the efficacy level). To prevent over-fitting, we used ridge regularization with a regularization strength of 1.

To measure the impact of the controllability representation in the networks, we interrupted the efficacy prediction to evaluate its impact on network performance in the

APE networks. We performed perturbations of the population representation of controllability. For this, we inverted the linear readout weights going from the population representation to the readout node. Given recurrent hidden layer activity h, an efficacy-readout value $\hat{\xi}'$, and a target efficacy that represents a perturbation of the ground truth $\hat{\xi}'_{\text{target}} = \xi' + \xi'_{\text{perturbstion}}$, we inverted the readout weights $W_{h,\hat{\xi}}$ to calculate one possible population activity that would correspond to $\hat{\xi}'_{\text{target}}$ based on the formula:

$$h_{\text{perturbed}} = h + W_{h,\hat{\xi}}^{-1}\left(\hat{\xi}'_{\text{target}} - \xi'\right)$$

We then use this formula to perturb the population activity to fixed levels as we expose the APE-trained networks to a variety of efficacy settings. Note that networks trained using a $\xi'_{\text{perturbation}} = 0$ will behave slightly differently from an unperturbed neural network since it will receive an advance indication of the given efficacy level in its distributed controllability representation.

In an additional analysis, we mistrain networks to misinterpret environmental control cues. To manipulate the population-level representation of efficacy that emerges in the LSTM (detected through a decoding analysis and through ablations of the efficacy node), We selectively unfroze the weights from input to LSTM and from LSTM to efficacy readout of converged APE-trained networks and continued training the networks for another 100k episodes on false APEs generated from binomial distributions corresponding to fixed probabilities.

Analysis of human behavior

We computed statistics for pairwise comparisons using Tukey's Honest Significant Differences test in the statsmodels package for Python (Seabold and Perktold, 2010). We only used data from test episodes where controllability to evaluate performance in these sections, so that many comparisons were across different groups. In cases where the data was paired, we used SciPy (Virtanen et al., 2020) for relative t-tests and statsmodels for Holm-Bonferroni multiple comparisons correction.

In addition to statistical tests, we evaluate adaptation across efficacy levels using linear mixed-effect models (see Supplement for full results). For these analyses, we considered both group and controllability in the episode as fixed effects, and participant ID as a random effect. We fit the mixed-effects models using statsmodels (Seabold and Perktold, 2010).

To evaluate how likely the APE-trained and standard models were as generative models of human trajectories, we first compute the log likelihood of the human trajectories under their respective policies. As a descriptive measure, we compute the average likelihood by averaging the stepwise (non-log) likelihoods of the human trajectories across all trials and episodes completed by an individual participant. We then average across different instantiations within a model type and compute statistics across different individual human participants. To quantify the difference, we use random-effects analysis for Bayesian model selection (Rigoux et al. 2014 and Stephan et al. 2009)

based on the total log-likelihood computed across all trials and episodes for each individual, and then averaged across different model instantiations.

We used scientific computing package PanDas (McKinney et al,. 2010) for data management.

Analysis of individual differences

To evaluate individual differences in performance, we constructed canonical correlation matrices. We computed the canonical correlation matrices separately for each of the groups, and then computed the average correlation matrix of the two weighted by the number of participants in each group. As pre-registered, we take the partial correlation coefficient instead of the canonical correlation coefficient when comparing the number of observe actions in task 1 to the number of sleep actions in task 2. We fully partial the number of observe actions in task 2 here to account for possible interference with the number of sleep actions that could arise from the correlation in observe behavior between task 1 and task 2 that was observed in human participants.

We construct the simulated population of perturbed participants using the methods described above. For task 1, we draw the perturbation magnitude for each individual from a standard normal distribution $\xi'_{\text{perturbation}} \sim \mathcal{N}(0,1)$. For the trait-level simulations, we maintain the same perturbation level for each individual across efficacy levels; for the random simulations, we resample the perturbation from the same distribution for each episode. For task 2, we maintain the same perturbations but re-center the values so that $\xi'_{\text{perturbation}} \sim \mathcal{N}(0.8, 0.5)$ instead to account for different task statistics. The random perturbation simulations for this task transform the used perturbation levels from task 1 similarly but additionally shuffle their order so that there is no individual consistency across tasks. We simulate populations of 150 individuals to have a similarly-sized distribution as for our human participants, but repeat each episode 100 times to have more robust estimates of behavioral outcomes.

We compare the correlation matrices for the human and simulated participants using cosine similarity. For the cases where we are comparing autocorrelation matrices (featuring the same variable along both rows and columns), we base the cosine similarity on the upper triangle matrix excluding the long diagonal.

Finally, we constructed 95%-confidence intervals for the cosine similarity by bootstrapping. We created 100 samples of the population human participants by resampling with replacement. The 95%-confidence intervals are then given by the 0.025- and 0.975-quantiles.

Analysis of psychiatric transdiagnostics

We calculated the psychiatric transdiagnostic factors using the published procedure and fitted models from Hopkins et al. (2022). We computed the Pearson correlation coefficients between the different transdiagnostic factors using SciPy and computed lines of best fit using NumPy's polyfit between transdiagnostics (Harris et al. 2020). We used statsmodels to relate A/D and compulsivity scores to behavioral measures

separately for high and low controllability settings, $\xi \in \{0, 0.125, 0.25, 0.375\}$ and $\xi \in \{0.625, 0.75, 0.875, 1\}$, respectively.

We constructed matrix representations of the distribution of performance measures across controllability levels by sorting the data into noniles based on the transdiagnostic factors. We renormalized the data between 0 and 1 based on maximum and minimum values for each behavioral measure separately for each group, and then combined the matrices using an average weighted by the number of participants in each group.

We compared the distributions in human behavior with those from the neural network simulations. For the trait simulations, we used the same simulations as in the previous sections, and computed noniles based on the value of the perturbation. To better capture the effect introduced by sorting based on perturbation, we compared performance with that of neural networks with $\xi'_{\text{perturbation}} = 0$. We normalized the matrices for both the trait- and zero-perturbed networks between 0 and 1 before computing cosine similarity to be in a similar range as for humans. We then compute the excess predictive power of the trait-simulated networks by comparing the cosine similarity of their distribution matrices with that of the unperturbed networks by the formula $\text{Fraction Remaining Similarity Explained} = \frac{\text{Similarity}_{\text{Trait}} - \text{Similarity}_{\text{Zero}}}{1 - \text{Similarity}_{\text{Zero}}}$.

We again computed 95%-confidence intervals by bootstrapping over 100 samples and taking the 0.025- and 0.975-quantiles in cosine similarity over the resulting distribution.

Supplement 1: Additional results

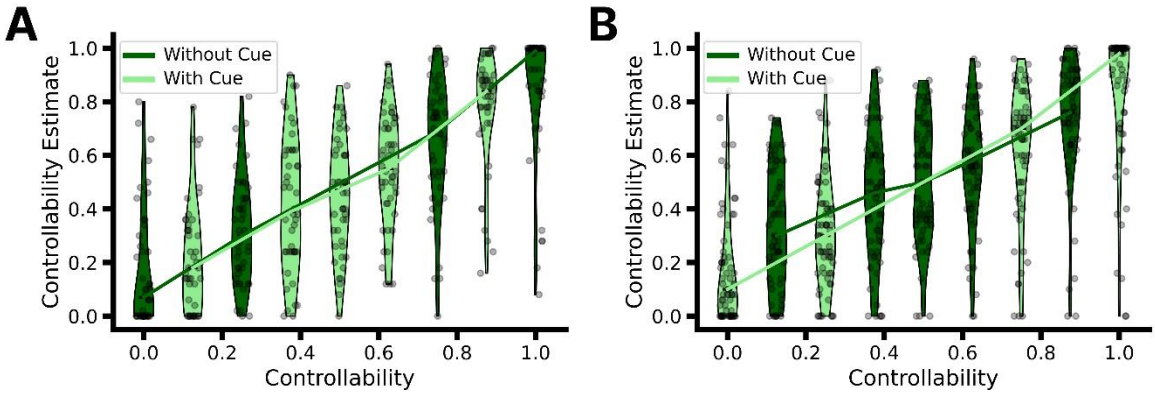

Fig. S1. Estimates of environmental controllability. On day 2, participants were asked to estimate the level of environmental controllability using a slider after each episode. (**A**) Estimates by participants in group A for both (*light green*) uncued training and (*dark green*) cued test episodes. (**B**) Same as (A), but for participants in group B.

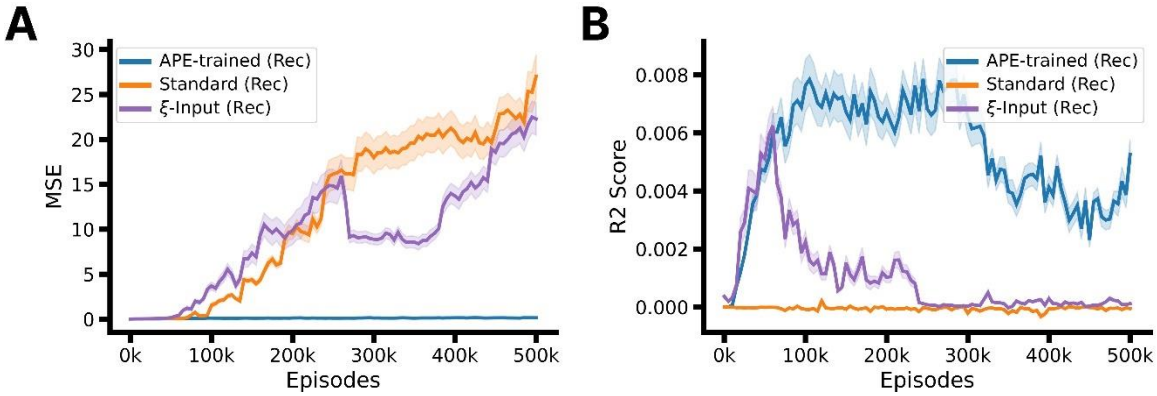

Fig. S2. Encoding of controllability. In addition to decoding controllability, we tested how strongly controllability was coded for in the LSTM layer of the networks using an encoding model. (**A**) Mean squared-error for the encoding model for (*blue*) APE-trained, (*orange*) standard, and (*purple*) ξ-input networks. (**B**) Same as (A), but using R2 score as the metric.

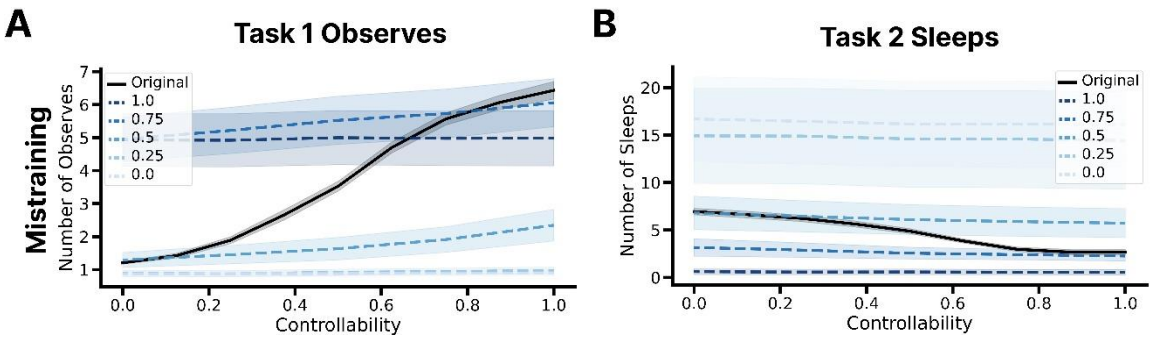

Fig. S3. Population mistraining. As an alternative to perturbations of the controllability representation, by selectively unfreezing the weights leading to the networks' controllability readout, we can retrain the network to believe it is operating under false

conditions by simulating fake action-prediction error sequences as the agent interacts with the environment. (**A**) In task 1, the networks sleep more across all controllability levels when (*darker*) they are mistrained to believe they are operating under conditions with higher efficacy, and the reverse for (*lighter*) networks trained to believe they are operating in low-efficacy conditions, compared with (*black*) the original neural network. (**B**) The same plot as (A) but for sleep actions on task 2, showing that networks mistrained to believe they have more control than they do sleep less, and vice versa. Note that the simulations depicted here use different model seeds as in the main text.

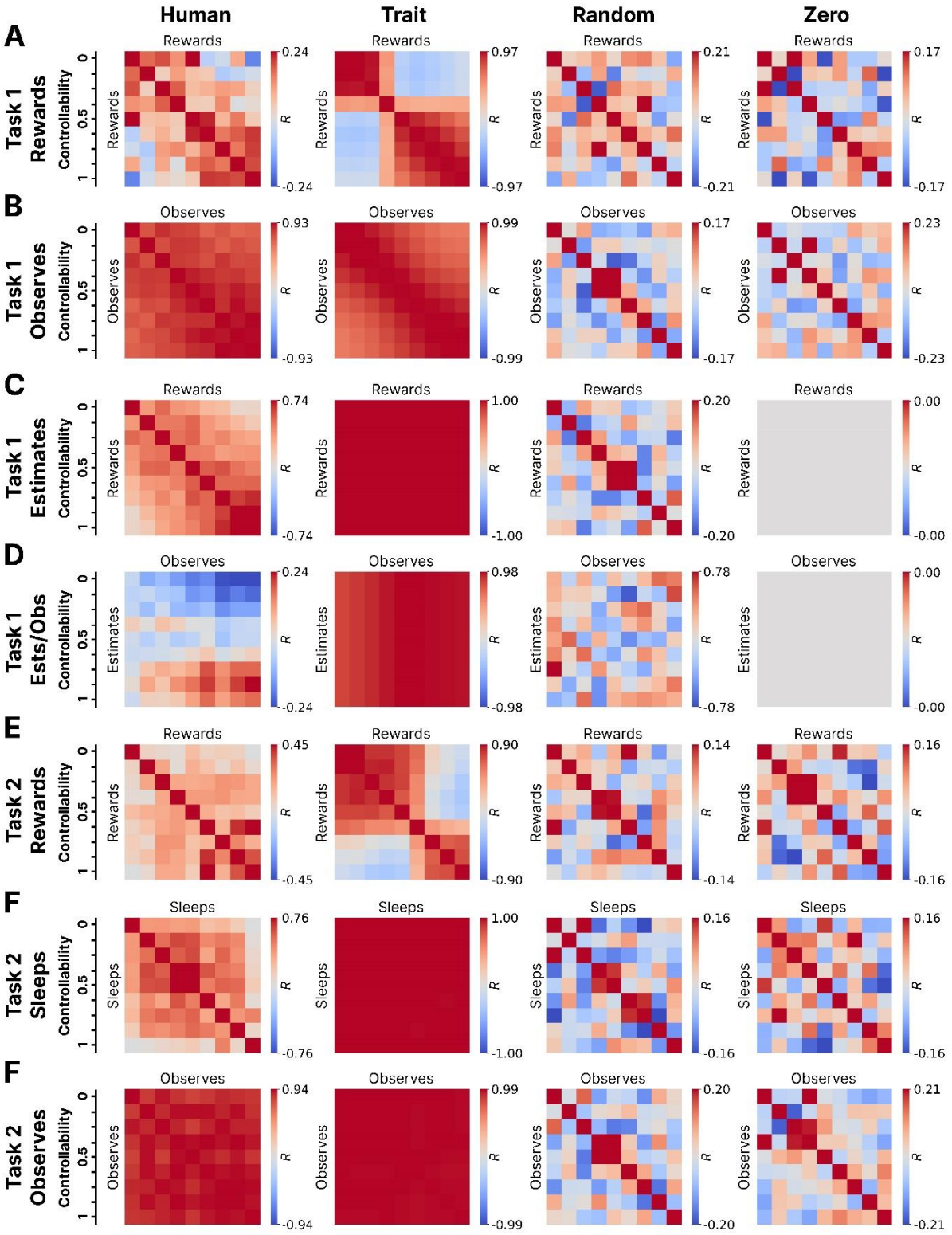

Fig. S4. Supplemental individual differences (Part 1). *See next page for legend.*

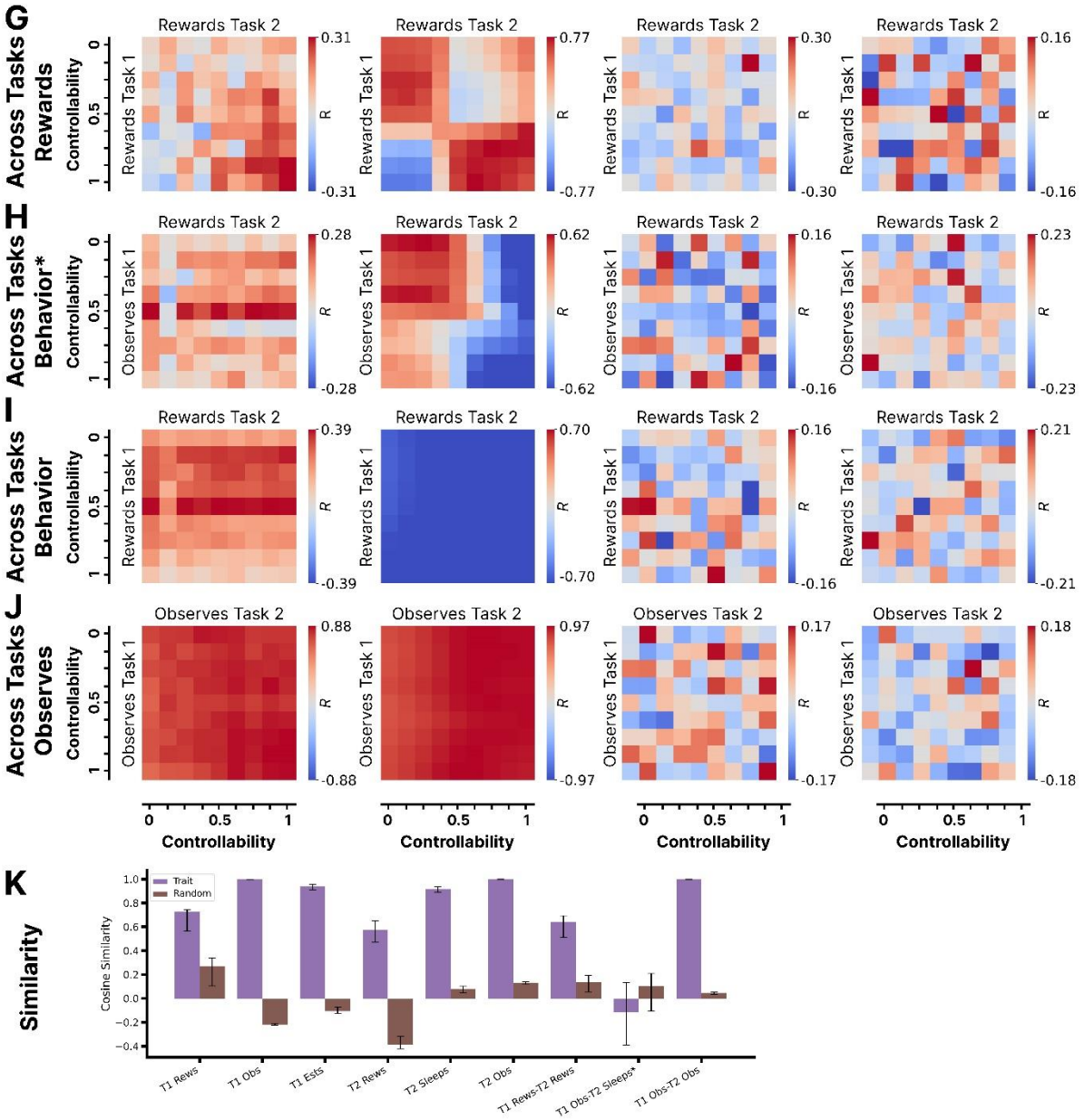

Fig. S5. Supplemental individual differences (Part 2). (**A-J**) Correlation structure for all behavioural variables for (*far left*) humans and (*middle left*) trait-perturbed neural networks, including (*middle right*) randomly perturbed neural networks and (*far right*) zero-perturbed neural networks. (**K**) Cosine similarity between the humans and both trait-simulated and randomly-perturbed neural networks along with bootstrapped 95%-confidence intervals.

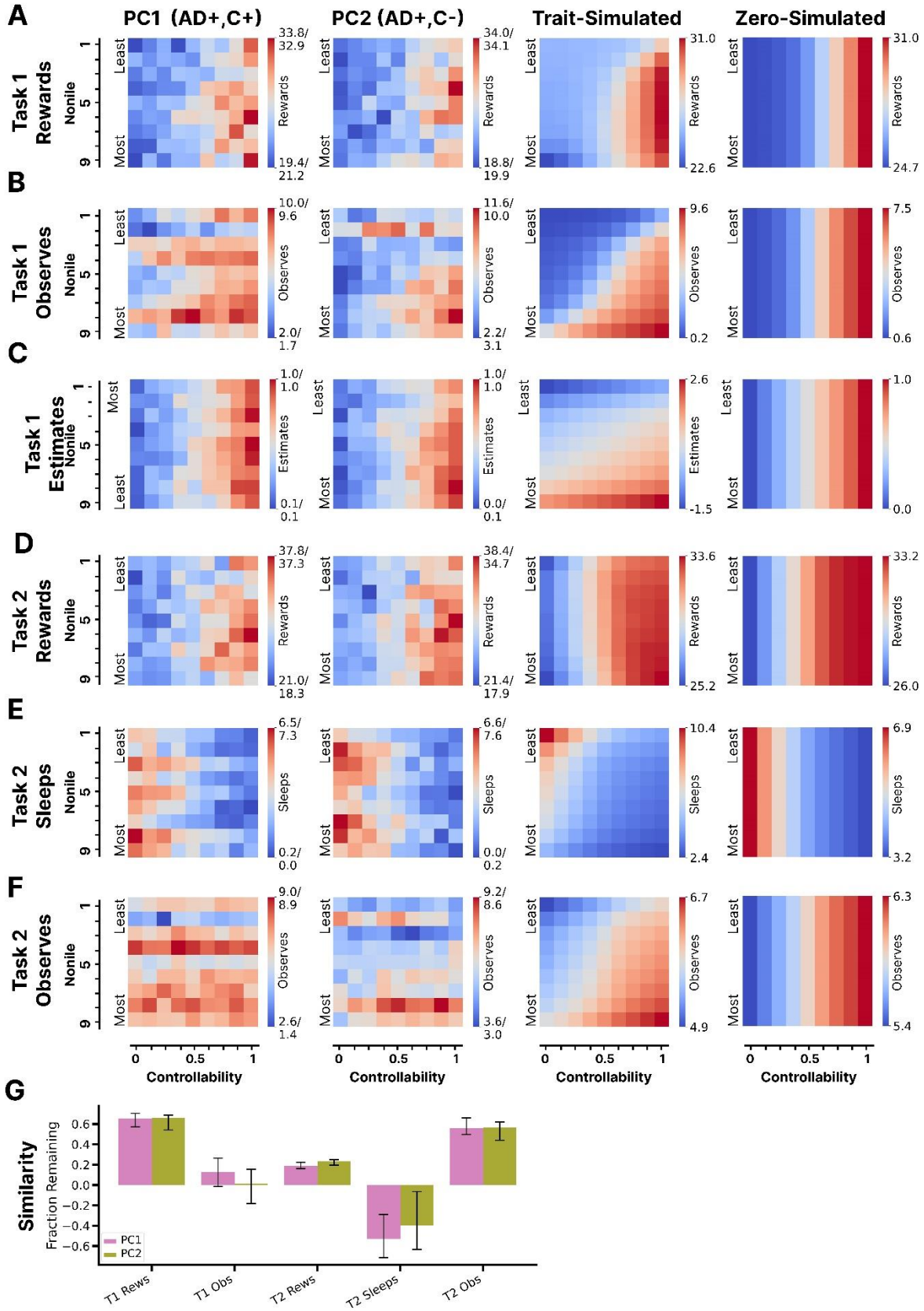

Fig. S6. Correlations with behavior for both PCs of transdiagnostic factors. (**A-F**) Heatmaps showing the interaction effect between transdiagnostics and efficacy on key behavioral variables for both PCs. (**G**) Cosine similarity for trait-simulated and zeros-simulated networks for the different transdiagnostic factors as depicted in **Figure** 6.

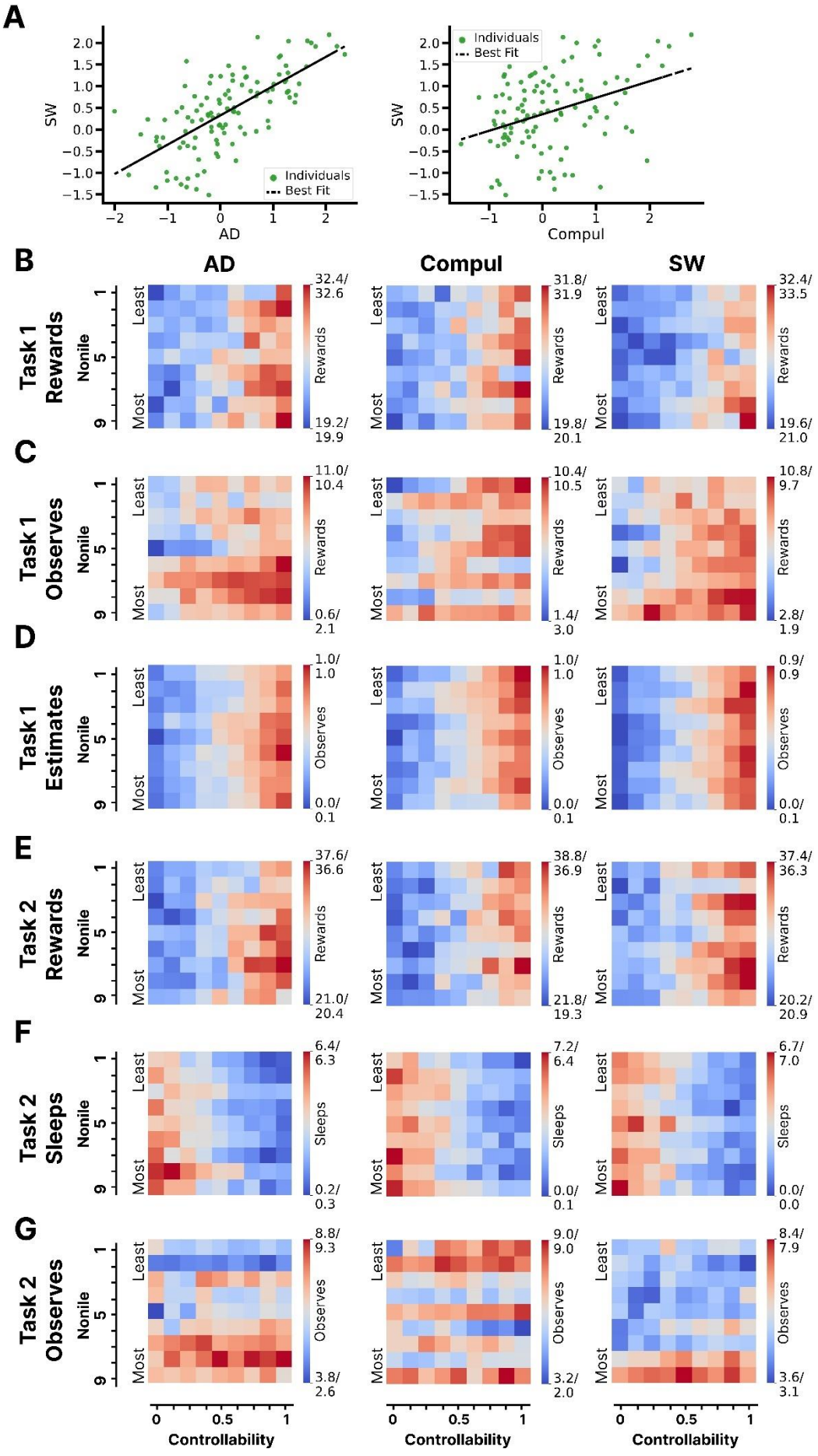
A
SW
AD
Compul
Individuals
Best Fit
B
C
D
E
F
G
AD
Compul
SW
Task 1 Rewards
Task 1 Observes
Task 1 Estimates
Task 2 Rewards
Task 2 Sleeps
Task 2 Observes
Nonile
Least
Most
Rewards
Observes
Sleeps
Controllability

Fig. S7. Supplemental transdiagnostics. (**A**) Scatter plot showing relationship between social withdrawal and (*left*) AD, (*right*) Compul. (**B-H)** Heatmaps showing the interaction effect between transdiagnostics and efficacy on key behavioral variables for both PCs.

Supplement 2: Full statistics

Performance improvements

Performance increased slightly from day 1 to day 2 for rewards aggregated across the whole dataset (total rewards: 224.05 ± 1.51, N=182, mean ± SEM for day 1; 226.64 ± 1.42, N=182, mean ± SEM for day 2; t(181) = -1.245, p = 0.107 according to an unpaired one-sided t-test, see **Error! Reference source not found.** for statistics for individual controllability settings). Despite the fact that the fact that the test task was harder because of the additional uncertainty due to the controllability, participants improve performance insignificantly from train to test set on both day 2 and day 3, potentially due to increases in skill (see **Error! Reference source not found.**).

Table S 1. Performance improvements from day 1 to day 2 evaluated on the test set using unpaired one-sided t-tests

Controll-ability	Mean Day 1	StdErr Day 1	Mean Day 2	StdErr Day 2	t-statistic	DoF	p-value	p-value corrected	Reject H0
0.000	23.177	0.410	22.988	0.430	0.316	180	0.624	1.000	False
0.125	23.674	0.472	23.688	0.370	-0.022	180	0.491	1.000	False
0.250	23.344	0.402	23.860	0.456	-0.849	180	0.199	1.000	False
0.375	24.488	0.488	24.646	0.400	-0.250	180	0.401	1.000	False
0.500	24.640	0.524	24.844	0.468	-0.290	180	0.386	1.000	False
0.625	25.151	0.445	26.052	0.477	-1.364	180	0.087	0.785	False
0.750	26.635	0.550	27.093	0.528	-0.593	180	0.277	1.000	False
0.875	27.744	0.577	26.896	0.658	0.955	180	0.829	1.000	False
1.000	26.750	0.761	27.337	0.726	-0.552	180	0.291	1.000	False

Table S 2. Performance improvements from train to test set using an unpaired two-sided t-tests

Day	Eff Set	Mean Train	StdErr Train	Mean Test	StdErr Test	DoF	T-statistic	P-value	Corrected p-value	Reject
2	[0.125, 0.375, 0.5, 0.625, 0.875]	25.021	0.275	25.225	0.242	180	-0.557	0.578	1.0	False

2	[0, 0.25, 0.75, 1.0]	25.188	0.270	25.320	0.310	180	-0.322	0.748	1.0 False
3	[0.125, 0.375, 0.5, 0.625, 0.875]	25.021	0.275	25.225	0.242	180	-0.557	0.578	1.0 False
3	[0, 0.25, 0.75, 1.0]	25.188	0.270	25.320	0.310	180	-0.322	0.748	1.0 False

Controllability estimates in task 1

Table S 3. Results of linear model for task 1

Model:	**MixedLM**	**Dependent Variable:**	**efficacy_estimates**
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	0.0484
Min. group size:	9	Log-Likelihood:	85.9446
Max. group size:	9	Converged:	Yes
Mean group size:	**9.0**		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-0.077	0.015	-5.050	0.000	-0.107	-0.047
C(group_C)[T.True]	-0.005	0.014	-0.331	0.740	-0.032	0.023
efficacy	0.280	0.022	12.973	0.000	0.238	0.322
Group Var	0.000	0.005				

Human observations in task 1

Table S 4: Tukey's HSDs

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower  upper  reject
-------------------------------------------------
   0.0    0.5   0.0366 0.0451 0.0006 0.0726   True
   0.0    1.0   0.0739    0.0  0.035 0.1129   True
   0.5    1.0   0.0373 0.0405 0.0013 0.0733   True
-------------------------------------------------
```

Table S 5: Linear model

Model:	**MixedLM**	**Dependent Variable:**	**n_observes**
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	5.9580

Min. group size:	9	Log-Likelihood:	-2458.3631
Max. group size:	9	Converged:	Yes
Mean group size:	**9.0**		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	5.082	0.484	10.506	0.000	4.134	6.031
C(group_C)[T.True]	-0.319	0.613	-0.521	0.603	-1.519	0.882
efficacy	2.777	0.239	11.598	0.000	2.308	3.246
Group Var	9.443	0.594				

Human rewards in task 1

Table S 6. Tukey's HSD for proportion of bets intended to place on correct light

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj lower  upper  reject
-------------------------------------------------
   0.0    0.5   0.0879 0.004 0.0239  0.152   True
   0.0    1.0   0.2294   0.0   0.16 0.2987   True
   0.5    1.0   0.1414   0.0 0.0774 0.2055   True
```

Table S 7. Tukey's HSD for rewards earned

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower  upper  reject
-------------------------------------------------
   0.0    0.5   2.1833 0.0899 -0.2587 4.6252  False
   0.0    1.0   6.8913    0.0  4.2486  9.534   True
   0.5    1.0    4.708    0.0  2.2661   7.15   True
-------------------------------------------------
```

Table S 8: Linear model for number of rewards earned

Model:	**MixedLM**	**Dependent Variable:**	**rewards_tallies**
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	19.9434
Min. group size:	9	Log-Likelihood:	-2948.6878
Max. group size:	**9**	Converged:	Yes
Mean group size:	**9.0**		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	5.082	0.484	10.506	0.000	4.134	6.031
C(group_C)[T.True]	-0.319	0.613	-0.521	0.603	-1.519	0.882
efficacy	2.777	0.239	11.598	0.000	2.308	3.246
Group Var	9.444	0.595				

Table S 9. Likelihood of APE vs. standard models for displayed controllability levels

Controllability	N	Mean APE	SEM APE	Mean Standard	SEM Standard
1.0	46	0.648	0.033	0.519	0.018
0.5	65	0.610	0.024	0.481	0.014
0.0	46	0.562	0.022	0.534	0.021

Table S 10. Likelihood of APE vs. standard models for all efficacy levels

Controllability	N	Mean APE	StdErr APE	Mean Standard	StdErr Standard
0.000	46	0.562	0.022	0.534	0.021
0.125	65	0.591	0.017	0.493	0.015
0.250	46	0.654	0.021	0.527	0.019
0.375	65	0.613	0.021	0.480	0.012
0.500	65	0.610	0.024	0.481	0.014
0.625	65	0.611	0.027	0.485	0.013
0.750	46	0.660	0.029	0.488	0.014
0.875	65	0.592	0.031	0.488	0.012
1.000	46	0.648	0.033	0.519	0.018

Human sleep actions in task 2

Table S 11. Tukey's HSD for human sleep rate for steps 8-20

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
==================================================
group1 group2 meandiff p-adj   lower   upper  reject
--------------------------------------------------
   0.0    0.5  -0.0841    0.0 -0.1271  -0.041   True
   0.0    1.0  -0.1286    0.0 -0.1752 -0.0821   True
   0.5    1.0  -0.0446 0.0404 -0.0876 -0.0015   True
--------------------------------------------------
```

Table S 12: Linear model for number of sleep actions

Model:	MixedLM	Dependent Variable:	n_sleeps
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	2.8097
Min. group size:	9	Log-Likelihood:	-2059.7559
Max. group size:	9	Converged:	Yes
Mean group size:	**9.0**		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	4.628	0.269	17.225	0.000	4.101	5.155
C(group_C)[T.True]	-0.302	0.334	-0.904	0.366	-0.957	0.353

efficacy	-3.575	0.164	-21.755	0.000	-3.897	-3.253
Group Var	2.698	0.258				

Human rewards in task 2

Table S 13. Tukey's HSD for number of rewards

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower  upper  reject
-------------------------------------------------
   0.0    0.5    2.686 0.0238 0.2902 5.0817   True
   0.0    1.0   9.8043    0.0 7.2117 12.397   True
   0.5    1.0   7.1184    0.0 4.7227 9.5141   True
-------------------------------------------------
```

Table S 14: Linear model for number of rewards

Model:	**MixedLM**	**Dependent Variable:**	**rewards_tallies**
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	32.2287
Min. group size:	9	Log-Likelihood:	-3195.7374
Max. group size:	9	Converged:	Yes
Mean group size:	**9.0**		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	24.148	0.503	47.964	0.000	23.161	25.135
C(group_C)[T.True]	-1.376	0.548	-2.509	0.012	-2.451	-0.301
efficacy	9.516	0.557	17.099	0.000	8.425	10.607
Group Var	4.517	0.212				

Human observe actions in task 2

Table S 15. Tukey's HSD for human observes in task 2

```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower  upper  reject
-------------------------------------------------
   0.0    0.5  -0.0077 0.8583  -0.042 0.0267  False
   0.0    1.0   0.0031 0.9787 -0.0341 0.0403  False
   0.5    1.0   0.0108 0.7397 -0.0236 0.0451  False
-------------------------------------------------
```

Table S 16. Linear model for number of observes in task 2

Model:	**MixedLM**	**Dependent Variable:**	**n_observes**
No. Observations:	999	Method:	REML
No. Groups:	111	Scale:	2.5760
Min. group size:	9	Log-Likelihood:	-2082.5541
Max. group size:	9	Converged:	Yes

Mean group size:	9.0					
	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	6.110	0.456	13.405	0.000	5.217	7.004
C(group_C)[T.True]	-0.824	0.587	-1.405	0.160	-1.974	0.326
efficacy	0.321	0.157	2.038	0.042	0.012	0.629
Group Var	4.517	0.212				

Table S 17. Likelihood of APE vs. standard models for task 2 for displayed controllability levels

Controllability	N	Mean APE	SEM APE	Mean Standard	SEM Standard
1.0	46	0.661	0.030	0.594	0.024
0.5	65	0.554	0.031	0.510	0.025
0.0	46	0.594	0.028	0.535	0.024

Table S 18. Likelihood of APE vs. standard models for task 2 for all controllability

Controllability	N	Mean APE	StdErr APE	Mean Standard	StdErr Standard
0.000	46	0.594	0.028	0.535	0.024
0.125	65	0.537	0.029	0.501	0.024
0.250	46	0.606	0.031	0.546	0.027
0.375	65	0.565	0.030	0.525	0.025
0.500	65	0.554	0.031	0.510	0.025
0.625	65	0.573	0.032	0.523	0.026
0.750	46	0.664	0.029	0.607	0.024
0.875	65	0.576	0.032	0.535	0.024
1.000	46	0.661	0.030	0.594	0.024

Individual differences

Table S 19. Cosine similarities with trait- and randomly-perturbed networks with bootstrapped confidence intervals

Task	Item	Comparison	Cosine Similarity	Lower	Upper	Name
0	T1	Rews	Trait	0.544	0.375	0.604
1	T1	Rews	Random	0.192	0.055	0.296
2	T1	Obs	Trait	0.993	0.977	0.996

3	T1	Obs	Random	-0.207	-0.224	-0.181
4	T1	Ests	Trait	0.876	0.811	0.911
5	T1	Ests	Random	-0.049	-0.120	0.019
6	T2	Rews	Trait	0.408	0.268	0.476
7	T2	Rews	Random	-0.174	-0.258	-0.038
8	T2	Sleeps	Trait	0.957	0.936	0.968
9	T2	Sleeps	Random	-0.041	-0.062	-0.014
10	T2	Obs	Trait	0.996	0.989	0.997
11	T2	Obs	Random	0.135	0.118	0.147
12	Across Tasks	T1 Rews-T2 Rews	Trait	0.444	0.305	0.503
13	Across Tasks	T1 Rews-T2 Rews	Random	0.206	0.084	0.292
14	Across Tasks	T1 Obs-T2 Sleeps*	Trait	0.222	-0.089	0.392
15	Across Tasks	T1 Obs-T2 Sleeps*	Random	0.166	0.012	0.245

Table S 20. Competitive linear regression for population observe correlation structure featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.488
Model:	OLS	Adj. R-squared:	0.440
Method:	Least Squares	F-statistic:	10.18
Date:	Sun, 02 Jun 2024	Prob (F-statistic):	7.36e-05
Time:	15:12:59	Log-Likelihood:	16.094
No. Observations:	36	AIC:	-24.19
Df Residuals:	32	BIC:	-17.85
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7489	0.075	9.938	0.000	0.595	0.902
x1	0.2383	0.045	5.289	0.000	0.147	0.330
x2	0.2754	0.379	0.727	0.472	-0.496	1.047
x3	0.3911	0.322	1.215	0.233	-0.265	1.047

Omnibus:	0.794	Durbin-Watson:	1.194
Prob(Omnibus):	0.672	Jarque-Bera (JB):	0.219
Skew:	-0.157	Prob(JB):	0.896
Kurtosis:	3.219	Cond. No.	27.3

```
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Table S 21. Competitive linear regression for population rewards correlation structure featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.385
Model:                            OLS   Adj. R-squared:                  0.327
Method:                 Least Squares   F-statistic:                     6.675
Date:                Sun, 02 Jun 2024   Prob (F-statistic):            0.00126
Time:                        15:18:16   Log-Likelihood:                 42.194
No. Observations:                  36   AIC:                            -76.39
Df Residuals:                      32   BIC:                            -70.05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0432      0.017      2.595      0.014       0.009       0.077
x1             0.0737      0.017      4.259      0.000       0.038       0.109
x2             0.2061      0.146      1.415      0.167      -0.091       0.503
x3             0.0145      0.156      0.093      0.926      -0.303       0.332
==============================================================================
Omnibus:                        1.751   Durbin-Watson:                   1.990
Prob(Omnibus):                  0.417   Jarque-Bera (JB):                0.768
Skew:                           0.035   Prob(JB):                        0.681
Kurtosis:                       3.712   Cond. No.                         14.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Table S 22. Competitive linear regression for population sleeps correlation structure for task 2 featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.140
Model:                            OLS   Adj. R-squared:                  0.059
Method:                 Least Squares   F-statistic:                     1.730
Date:                Sun, 02 Jun 2024   Prob (F-statistic):              0.181
Time:                        15:21:21   Log-Likelihood:                5.1735
No. Observations:                  36   AIC:                            -2.347
Df Residuals:                      32   BIC:                             3.987
Df Model:                           3
Covariance Type:            nonrobust
```

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4242	0.287	1.475	0.150	-0.161	1.010
x1	0.0120	0.080	0.149	0.883	-0.152	0.176
x2	0.5250	0.411	1.279	0.210	-0.311	1.362
x3	0.8627	0.464	1.858	0.072	-0.083	1.809

Omnibus:	0.736	Durbin-Watson:	1.880
Prob(Omnibus):	0.692	Jarque-Bera (JB):	0.613
Skew:	-0.303	Prob(JB):	0.736
Kurtosis:	2.798	Cond. No.	47.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table S 23. Competitive linear regression for population observes correlation structure for task 2 featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.169
Model:	OLS	Adj. R-squared:	0.091
Method:	Least Squares	F-statistic:	2.166
Date:	Sun, 02 Jun 2024	Prob (F-statistic):	0.111
Time:	15:25:26	Log-Likelihood:	15.444
No. Observations:	36	AIC:	-22.89
Df Residuals:	32	BIC:	-16.55
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2626	0.565	4.006	0.000	1.112	3.413
x1	-0.3555	0.239	-1.485	0.147	-0.843	0.132
x2	0.5584	0.302	1.851	0.073	-0.056	1.173
x3	0.0213	0.334	0.064	0.949	-0.658	0.701

Omnibus:	0.327	Durbin-Watson:	1.332
Prob(Omnibus):	0.849	Jarque-Bera (JB):	0.470
Skew:	-0.187	Prob(JB):	0.790
Kurtosis:	2.583	Cond. No.	56.8

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table S 24. Competitive linear regression for population rewards correlation structure for task 2 featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.093
Method:	Least Squares	F-statistic:	0.008289
Date:	Sun, 02 Jun 2024	Prob (F-statistic):	0.999
Time:	15:28:12	Log-Likelihood:	29.514
No. Observations:	36	AIC:	-51.03
Df Residuals:	32	BIC:	-44.69
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1645	0.025	6.542	0.000	0.113	0.216
x1	-0.0025	0.036	-0.069	0.945	-0.075	0.070
x2	-0.0277	0.263	-0.105	0.917	-0.563	0.507
x3	0.0306	0.247	0.124	0.902	-0.473	0.534

Omnibus:	7.850	Durbin-Watson:	0.828
Prob(Omnibus):	0.020	Jarque-Bera (JB):	6.430
Skew:	0.924	Prob(JB):	0.0402
Kurtosis:	3.933	Cond. No.	16.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table S 25. Competitive linear regression for population correlation structure for observes task 1 and sleeps task 2 (with observes task 2 partialed out) featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.013
Model:	OLS	Adj. R-squared:	-0.025
Method:	Least Squares	F-statistic:	0.3455
Date:	Sun, 02 Jun 2024	Prob (F-statistic):	0.792
Time:	15:30:56	Log-Likelihood:	95.492
No. Observations:	81	AIC:	-183.0
Df Residuals:	77	BIC:	-173.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0665	0.009	7.494	0.000	0.049	0.084
x1	-0.0012	0.013	-0.100	0.921	-0.026	0.024
x2	0.0988	0.113	0.875	0.384	-0.126	0.324
x3	-0.0587	0.100	-0.584	0.561	-0.259	0.141

Omnibus:	0.786	Durbin-Watson:	1.384

```
Prob(Omnibus):                  0.675   Jarque-Bera (JB):                0.376
Skew:                          -0.139   Prob(JB):                        0.829
Kurtosis:                       3.185   Cond. No.                         13.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Table S 26. Competitive linear regression for population rewards correlation structure for across tasks featuring the trait-simulated (x1), random perturbation (x2), and zero-perturbation models (x3)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.171
Model:                            OLS   Adj. R-squared:                  0.139
Method:                 Least Squares   F-statistic:                     5.310
Date:                Sun, 02 Jun 2024   Prob (F-statistic):            0.00222
Time:                        15:34:05   Log-Likelihood:                 69.723
No. Observations:                  81   AIC:                            -131.4
Df Residuals:                      77   BIC:                            -121.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0625      0.015      4.171      0.000       0.033       0.092
x1             0.0962      0.026      3.751      0.000       0.045       0.147
x2            -0.0979      0.148     -0.661      0.511      -0.393       0.197
x3             0.0818      0.138      0.593      0.555      -0.193       0.357
==============================================================================
Omnibus:                        1.533   Durbin-Watson:                   1.130
Prob(Omnibus):                  0.465   Jarque-Bera (JB):                1.434
Skew:                          -0.206   Prob(JB):                        0.488
Kurtosis:                       2.495   Cond. No.                         14.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

Transdiagnostic factors

Table S 27. Cosine similarities with trait- and unperturbed networks with bootstrapped 95%-confidence intervals for the first two principal components fitted to A/D and compulsivity scores

Task	Transdiagnostic	item	Comparison	Cosine Similarity	Lower	Upper
T1	PC1	Rews	Trait	0.963	0.943	0.965

T1	PC1	Rews	Zeros	0.893	0.858	0.898
T1	PC2	Rews	Trait	0.957	0.937	0.958
T1	PC2	Rews	Zeros	0.874	0.839	0.881
T1	PC1	Obs	Trait	0.853	0.829	0.877
T1	PC1	Obs	Zeros	0.831	0.809	0.841
T1	PC2	Obs	Trait	0.853	0.812	0.863
T1	PC2	Obs	Zeros	0.850	0.823	0.853
T1	PC1	Ests	Trait	0.880	0.872	0.889
T1	PC1	Ests	Zeros	0.984	0.974	0.984
T1	PC2	Ests	Trait	0.892	0.877	0.902
T1	PC2	Ests	Zeros	0.981	0.971	0.982
T2	PC1	Rews	Trait	0.963	0.953	0.965
T2	PC1	Rews	Zeros	0.955	0.942	0.955
T2	PC2	Rews	Trait	0.970	0.956	0.969
T2	PC2	Rews	Zeros	0.961	0.944	0.959
T2	PC1	Sleeps	Trait	0.859	0.823	0.891
T2	PC1	Sleeps	Zeros	0.908	0.886	0.914
T2	PC2	Sleeps	Trait	0.889	0.856	0.919
T2	PC2	Sleeps	Zeros	0.920	0.900	0.928
T2	PC1	Obs	Trait	0.921	0.903	0.940
T2	PC1	Obs	Zeros	0.822	0.800	0.840
T2	PC2	Obs	Trait	0.916	0.877	0.932
T2	PC2	Obs	Zeros	0.805	0.783	0.830

Table S 28. Cosine similarities with trait- and unperturbed networks with bootstrapped confidence intervals on original transdiagnostic factors

Task	Transdiagnostic	Item	Comparison	Cosine Similarity	Lower	Upper
T2	Compul	Rews	Trait	0.968	0.957	0.970
T2	Compul	Rews	Zeros	0.963	0.949	0.963
T2	AD	Rews	Trait	0.968	0.955	0.970
T2	AD	Rews	Zeros	0.961	0.946	0.962
T2	SW	Rews	Trait	0.969	0.954	0.970

T2	SW	Rews	Zeros	0.963	0.943	0.961
T2	Compul	Sleeps	Trait	0.887	0.851	0.896
T2	Compul	Sleeps	Zeros	0.923	0.895	0.924
T2	AD	Sleeps	Trait	0.860	0.834	0.888
T2	AD	Sleeps	Zeros	0.923	0.904	0.927
T2	SW	Sleeps	Trait	0.908	0.877	0.925
T2	SW	Sleeps	Zeros	0.921	0.897	0.924
T2	Compul	Obs	Trait	0.883	0.866	0.912
T2	Compul	Obs	Zeros	0.820	0.794	0.838
T2	AD	Obs	Trait	0.921	0.900	0.936
T2	AD	Obs	Zeros	0.798	0.778	0.824
T2	SW	Obs	Trait	0.930	0.890	0.935
T2	SW	Obs	Zeros	0.820	0.786	0.831
T2	Compul	Rews	Trait	0.968	0.957	0.970
T2	Compul	Rews	Zeros	0.963	0.949	0.963
T2	AD	Rews	Trait	0.968	0.955	0.970
T2	AD	Rews	Zeros	0.961	0.946	0.962
T2	SW	Rews	Trait	0.969	0.954	0.970
T2	SW	Rews	Zeros	0.963	0.943	0.961
T2	Compul	Sleeps	Trait	0.887	0.851	0.896
T2	Compul	Sleeps	Zeros	0.923	0.895	0.924
T2	AD	Sleeps	Trait	0.860	0.834	0.888
T2	AD	Sleeps	Zeros	0.923	0.904	0.927
T2	SW	Sleeps	Trait	0.908	0.877	0.925
T2	SW	Sleeps	Zeros	0.921	0.897	0.924
T2	Compul	Obs	Trait	0.883	0.866	0.912
T2	Compul	Obs	Zeros	0.820	0.794	0.838
T2	AD	Obs	Trait	0.921	0.900	0.936
T2	AD	Obs	Zeros	0.798	0.778	0.824
T2	SW	Obs	Trait	0.930	0.890	0.935
T2	SW	Obs	Zeros	0.820	0.786	0.831
T2	Compul	Rews	Trait	0.968	0.957	0.970
T2	Compul	Rews	Zeros	0.963	0.949	0.963
T2	AD	Rews	Trait	0.968	0.955	0.970

T2	AD	Rews	Zeros	0.961	0.946	0.962
T2	SW	Rews	Trait	0.969	0.954	0.970
T2	SW	Rews	Zeros	0.963	0.943	0.961
T2	Compul	Sleeps	Trait	0.887	0.851	0.896
T2	Compul	Sleeps	Zeros	0.923	0.895	0.924
T2	AD	Sleeps	Trait	0.860	0.834	0.888

Supplement 3: Results of pilot study

The results presented in the main text are a replication of a pilot study conducted with n=80 participants (n=60 after the same exclusions registered in this task). Note that the version of task 2 used here uses a higher volatility value than the one used in the main text and therefore is not compared with the neural networks directly.

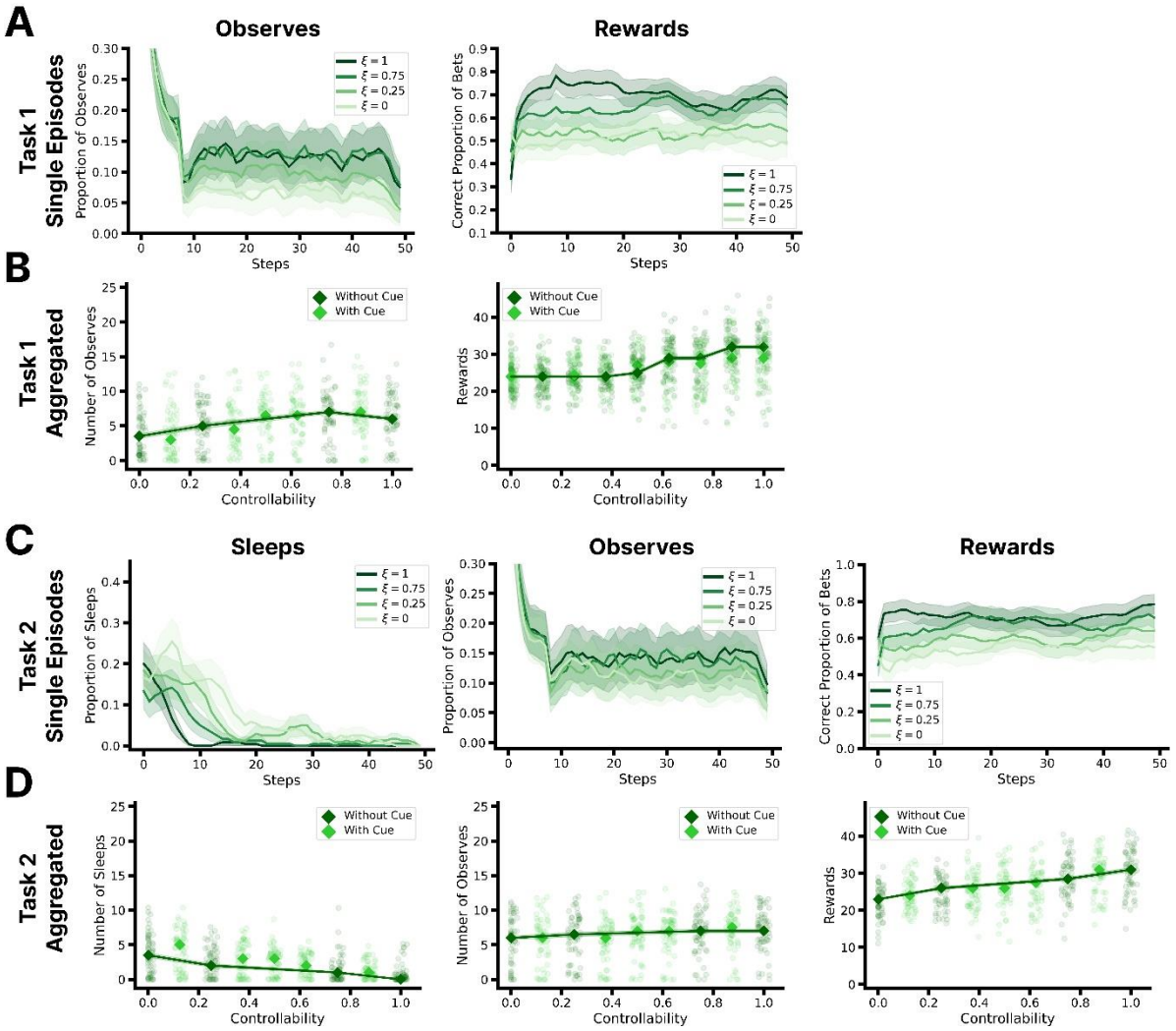

Fig. S8. Replication of task behavior. (**A**) Trial-wise behavior on task 1. (**B**) Variation of behavior across different levels of controllability for task 1. (**C**) Trial-wise behavior on task 2. (**D**) Variation of behavior across different levels of controllability for task 2.

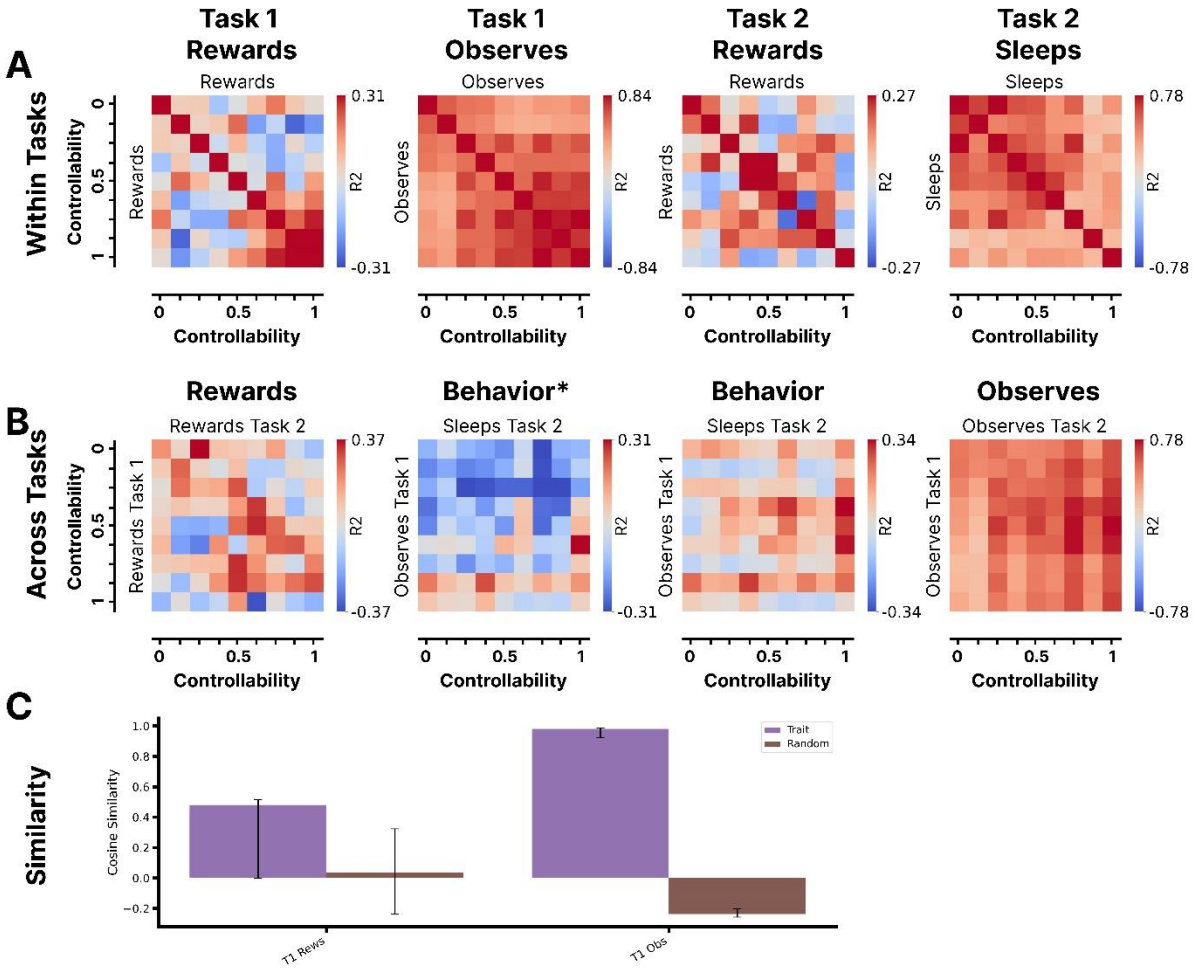

Fig. S9. Replication of individual differences. (**A-B**) Correlation in participant behavior across different controllability settings. The star indicates that we are partialling out number of observe actions on task 2. (**C**) Cosine similarity between human correlation matrices and the trait-perturbed and randomly-perturbed neural networks on task 1.

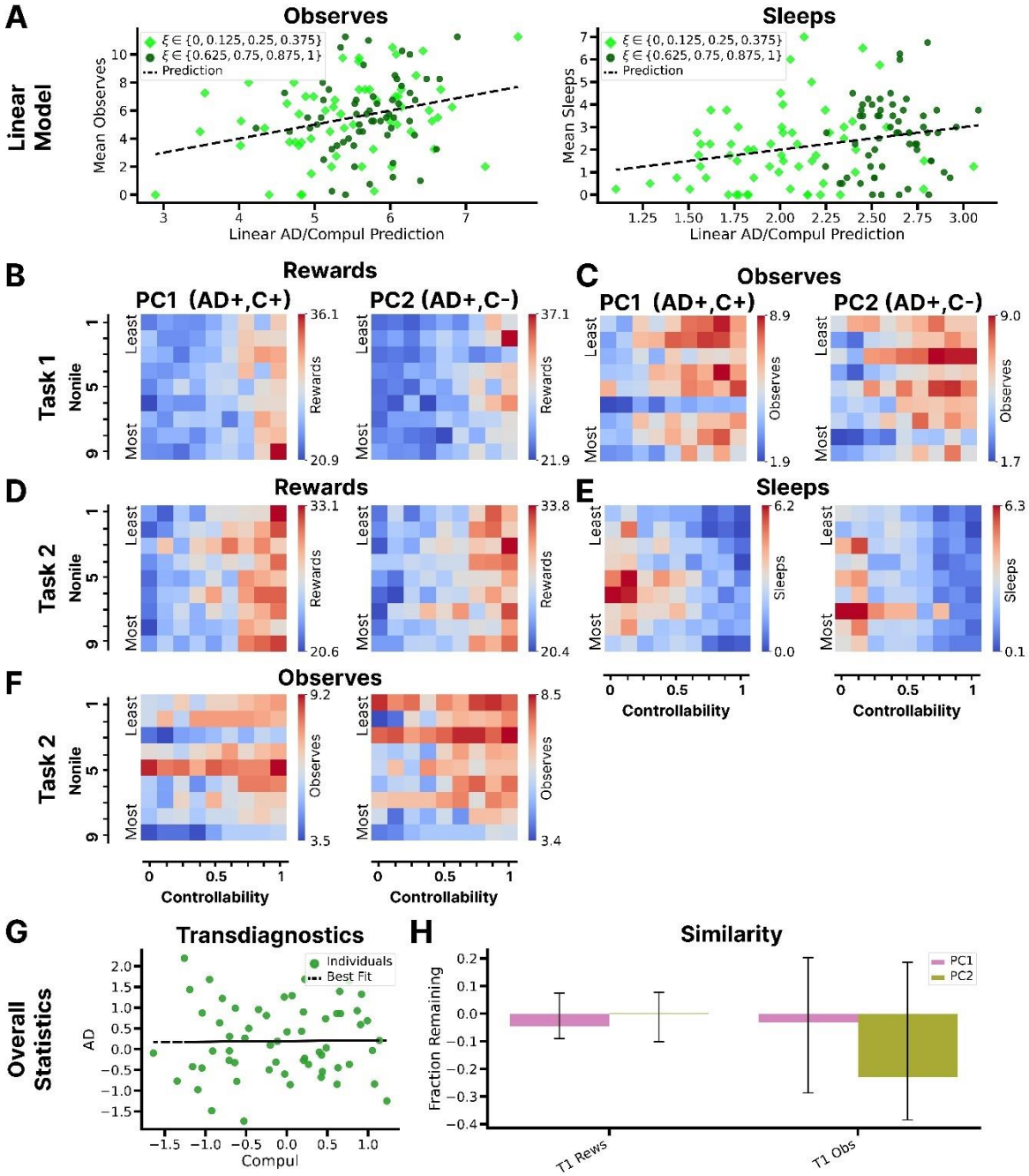

Fig. S10. Replication of transdiagnostics. (**A**) Predictions of linear models for number of observe and number of sleep actions based on transdiagnostic factors A/D and compulsivity. (**B-F**) Behavior as a function of controllability in the episode and participant transdiagnostic scores divided into noniles. (**G**) Correlation between A/D and compulsivity transdiagnostic scores. (**H**) Fraction of the remaining cosine similarity between humans and the unperturbed neural networks that is explained by the APE-based neural networks.