State-Independent and State-Dependent Learning in a Motivational Go/NoGo task

Azadeh Nazemorroaya[1,2,3](azadeh.nazemorroaya@tuebingen.mpg.de), Dan Bang[4,5] (danbang@cfin.au.dk), Peter Dayan[1,2](dayan@tue.mpg.de)

[1]Max Planck Institute for Biological Cybernetics, Tübingen, Germany; [2]University of Tübingen, Tübingen, Germany; [3]International Max Planck Research School for the Mechanisms of Mental Function and Dysfunction; [4]Center of Functionally Integrative Neuroscience, Aarhus University, Aarhus, Denmark; [5]Fralin Biomedical Research Institute at VTC, Virginia Tech, Roanoke, USA

Abstract

Recent research has identified substantial individual differences in how people solve value-based tasks. Here, we examine such differences in the motivational Go/NoGo task, which orthogonalizes action and valence, using open-source data from 817 participants. Using computational modeling and behavioral analysis, we identified four distinct clusters of people. Three clusters corresponded to previous models of the task, including people with different learning rates for cues that signal rewarding and punishing states and with different sensitives for rewards and punishments. The fourth cluster of people acted like naïve reinforcement learners, with their responses shaped by outcomes in a manner that was independent of the state information provided by the cues. In addition to providing evidence that state-independent learning is a common disposition, we show that not considering such learning can dramatically affect the results of computational modeling. We discuss the implications for the modeling of data from heterogeneous populations.

Keywords: motivational Go/NoGo task; value-based decision-making; state-independent learning

Introduction

Human behavior reflects a complex mixture of psychological forces. In the context of learning, particular conflict can arise between Pavlovian and instrumental responding. In the Pavlovian responding, value is automatically tied to action--so we are instinctively invigorated in favorable circumstances, but cautious in threatening ones (Dayan et al., 2006). In instrumental responding, by contrast, the choice of what to do is contingent on the consequences of the actions. (Dayan et al., 2006). Critically, not everyone is equally susceptible to the Pavlovian lure of, for instance, a smart outfit in a shopping center, especially when accompanied by pleasant music. While the attraction of such an outfit can be irresistible, for some, rational action selection might override value-seeking, depending on factors such as context and personality traits.

The motivational Go/NoGo task (Fig. 1) tests this relationship. It includes four distinct states wherein motivation can either align, or interfere, with appropriate action selection: press a button to win points (Go+), press a button to avoid losing points (Go-), not press a button to win points (NoGo+), and not press a button to avoid losing points (NoGo-). By crossing valence and action, the task can measure the extent to which Pavlovian biases can dominate over instrumentality. Studies have shown that population-level performance falls significantly short of its optimum, particularly in the Pavlovian-incongruent scenarios where one must act to avoid punishment (Go-) or refrain from acting to win (NoGo+) (Guitart-Masip et al., 2012; Moutoussis et al., 2018). However, individual performance on the motivational Go/NoGo task is known to vary substantially.

While not fully explored in the context of the motivational Go/NoGo task, studies suggest that a key source of such individual variability lies in the way value is processed and contributes to decision-making. For instance, studies have widely reported asymmetric learning (Michely et al., 2022) and differences in performance in rewarding versus punishing states of affectively charged tasks (Cavanagh et al., 2011; Kim et al., 2014). This variation correlates with individual cognitive states, as suggested by the work of Mkrtchian, Roiser & Robinson et al., (2017) and Proulx, Hikosaka, & Malinow et al., (2014). Another potential source of individual variation in value processing that has come to light more recently involves credit assignment. Although value should normally be assigned to the states and actions responsible for generating it, value can sometimes be attributed to irrelevant task features. For instance, in a 2-step probabilistic binary-choice task (Shahar et al., 2019), credit assignment to a right/left key was influenced by recent rewarding experiences from actions involving that key, even though the semantics of the keypress, as defined by state information, were different on each trial. A follow-up study showed that the extent of the influence of such state-independent learning correlated with measures of compulsivity (Shahar et al., 2021). This result is notable as compulsive behaviors may be founded on the (false) belief that engaging in a particular action can avert negative outcomes, even when there is no causal relationship with the events they aim to influence. In support of a hypothesis that overcoming state-independent learning requires cognitive control, another follow-up study found state-independent learning to be negatively correlated with working memory capacity (Ben-Artzi, Luria & Shahar, 2022). While these studies indicate that state-independent learning is a robust behavioral effect, and possibly even an individual trait, the hypothesis remains to be tested in other affectively charged tasks. In addition, an investigation of the impact of state-independent learning on computational modeling has yet to be conducted.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

In this study, we employed computational modeling to explore how individual differences in both state-dependent and state-independent learning are reflected in solutions to an affectively charged task. We analyzed published data from 817 participants performing the motivational Go/NoGo task (Guitart-Masip et al., 2012; Moutoussis et al., 2018) in which cues (which we refer to as 'states') were presented randomly with equal repetitions in general. Participants had to learn by trial-and-error whether (Go) or not (NoGo) to press a button to obtain a reward or avoid a punishment for a given cue. We used model fitting and comparison to show how behavior differs across this population. We identified four distinct clusters of participants, each approaching the task differently, as evidenced by the distinct behavioral profiles associated with each cluster. Two groups learnt the task successfully, with one group displaying the highest performance, while the other group showed high decision variability in punishing states leading to a lower performance in these states. The other two groups of participants performed rather poorly on the task, with 9% downplaying the reward contingencies of actions, and 18% downplaying the reward contingencies of cues, with the latter exhibiting a form of naïve state-independent learning. Furthermore, we demonstrated that, if state-independent learning is indeed a common behavioral disposition, then failing to accommodate it can have a dramatic impact on the results of computational modeling, by distorting the estimation of fitted parameters. This distortion would make them less reliable indicators of individual behavioral dispositions (e.g., Huys, Maia & Frank, 2016).

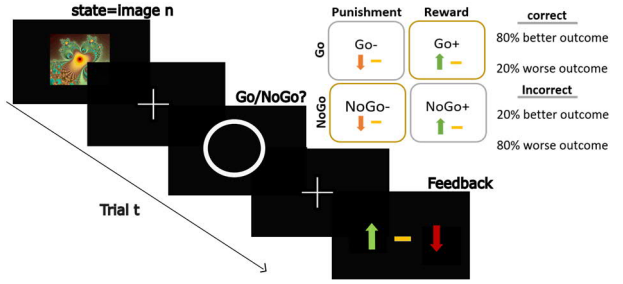

Fig. 1. Schematic of the motivational Go/NoGo task. The task features four distinct states: Go+ to win points, Go- to avoid losing points, NoGo+ to win points, and NoGo- to avoid losing points. Participants are required to learn to press a button (Go) or refrain from pressing a button (NoGo) in each state through trial and error. In potentially rewarding states, participants receive either a positive point, indicated by a green arrow, or neutral feedback, shown as a yellow rectangle. Conversely, in potentially punishing states, the feedback is either a negative point, represented by a red arrow, or neutral. Feedback is probabilistic such that even after a correct response, the better outcome for each state is delivered only 80% of the time. In two task states (Go+ and NoGo-), the correct response aligns with reflexes (Pavlovian-congruent), while in the other two task states (Go- and NoGo+), the correct response conflicts with reflexes (Pavlovian-incongruent).

Results

We analyzed open-source data from 817 healthy participants, aged between 14 to 24 years (Moutoussis et al., 2018) who performed the motivational Go/NoGo task (Moutoussis et al., 2018). For parameter estimation and model comparison, we utilized the Computational Behavioral Modeling (CBM) package, which is based on a Hierarchical Bayesian Inference algorithm (Piray et al., 2019). CBM has two objectives: 1) comparing the evidence supporting competing models and 2) estimating the free parameters within each of these models at both the individual and group levels. CBM incorporates a random-effect assumption in the space of models, which facilitates the analysis of individual differences. Specifically, it calculates a responsibility ratio (r_{kn}) which describes the probability of participant n's data being generated by model k; r_{kn} ranges between 0 and 1 for each participant n and model k pair, with a higher r_{kn} value indicating a greater likelihood that the data of participant n is generated by model k and the sum across all models k equaling 1 for each participant n. The r_{kn} value, which scales each participant's contribution to the group-level parameter estimation for a given model, makes model comparison straightforward: model comparison can be achieved by enumerating the responsibility ratios across the group in favor of each model and thereby computing normalized model frequencies across the population as well (Fig. 2A). In addition, by inspecting the maximum r_{kn} for each participant, CBM can be used to group participants within the model space and the corresponding group-level parameters for a given model can be used to understand the behavioral profile of each of these groups.

Our use of CBM enabled participant categorization, and identification of state-independent learners for model-agnostic analysis. In the process of refining our model space, we initially incorporated a range of models, either proposed by previous studies or based on our state-independent learning hypothesis (Guitart-Masip et al., 2012; Moutoussis et al., 2018): M, M2LR, M2invT and Msind. In addition, we included versions of the first three models which incorporated a state-independent learning component, M2invTsind, and M2LRsind. Through model fitting and comparison, we refined the model space to meet specific criteria: convergence within CBM (convergence defined by the change in normalized parameter values between two consecutive iterations being less than 0.01) and including only models applicable to a significant number of participants at this stage (N>20; according to reported CBM performance as a function of number of participants (Piray et al., 2019)). We then examined the convergence of CBM in the absence of any excluded models. Ultimately, given the evidence that: 1) model frequency of each model across population was significantly more than defined criteria; and 2) the behavioral structure of four identified groups are significantly different from each other, we defined a model space that includes four distinct models: a basic model M; model M2LR, which was the winning model in the original analysis of the data (Moutoussis et al., (2018); and a popular variation of model

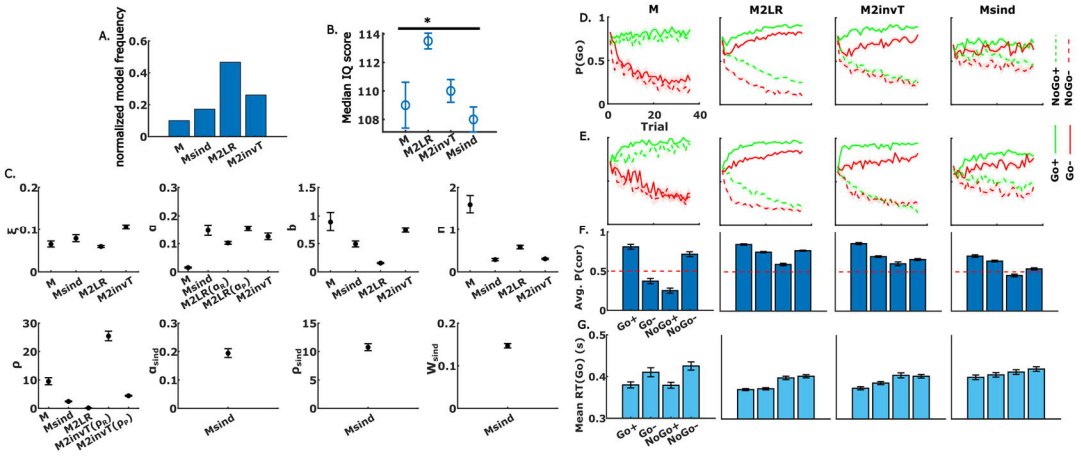

Fig. 2A. Model frequencies across 817 participants. B. Group-level median IQ (WASI) ± SEM for each cluster. C. Group-level mean parameters ± SEM for each cluster (where the parameters are present in the models concerned). D-G. Group-level mean data ± SEM for (D) the empirical probability of making a Go response, (E) the model-inferred probability of making a Go response, (F) response accuracy and (G) reaction times for Go responses shown separately for the four clusters (columns). The M2LR and M2invT clusters represent more natural learning trajectories with (D-E) solid lines (P(Go) in Go+ and Go- states) increasing and dashed lines (P(Go) in NoGo+ and NoGo- states) decreasing gradually across trials and (G) participants making faster Go response in Go states than NoGo states. The M cluster relied heavily on state valence, (D-E) largely making Go and NoGo responses in rewarding and punishing states, respectively. The Msind cluster exhibited non-instrumental behavior, characterized by frequent switching of responses, as illustrated by (D-E) the nearly flat and close-to-chance learning trajectories.

M, M2invT. We included model Msind with a state-independent learning component to examine our hypothesis about the presence and extent of a win-stay-lose-shift strategy in this task (Table 1).

Model M is based on the use of the Rescorla-Wagner rule to learn state-values (V) and state-action values (Q) (eq. 1-4). This involves two parameters: a learning rate (α) and outcome sensitivity (ρ; equivalent to inverse temperature). The model also includes two further parameters that influence action propensities: a fixed Go bias (b; eq. 5), which captures any inclination to act rather than remaining passive and defines the starting point of the state-specific learning curves; and a Pavlovian bias (π; eq. 5), which, if positive, boosts Go over NoGo responses in states with positive values and NoGo responses over Go responses in states with negative values. Finally, actions are selected according to a SoftMax probabilistic policy (eq. 6), including a lapse parameter ξ.

Table1 models and parameters

Models	parameters
M	Learning rate (α), Sensitivity (ρ), Go bias (b), Pavlovian bias (π)
M2LR	M + two distinct learning rates for reward and punishment (α_R, α_P)
M2invT	M + two distinct sensitivities (equivalent to inverse temperature) for reward and punishment (ρ_R, ρ_P)
Msind	M+state-independent learning rate (α_{sind}), state-independent sensitivity (ρ_{sind}), state-independent weight (w_{sind})

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \alpha \times \delta_1(t) \quad \text{eq. 1}$$

$$\delta_1(t) = \rho \times outcome(t) - Q_{t-1}(a_t, s_t) \quad \text{eq. 2}$$

$$V_t(s_t) = V_{t-1}(s_t) + \alpha \times \delta_2(t) \quad \text{eq. 3}$$

$$\delta_2(t) = \rho \times outcome(t) - V_{t-1}(s_t) \quad \text{eq. 4}$$

$$W(t,a) = \begin{cases} Q_t(a_t, s_t) + b + \pi \times V(s_t) & if\ a_t = Go \\ Q_t(a_t, s_t) & if\ a_t = NoGo \end{cases} \quad \text{eq. 5}$$

$$P(a_t|s_t) = \frac{\exp(W(t,a_t))}{\exp(\sum_a W(t,a))} \times (1-\xi) + \frac{\xi}{2} \quad \text{eq. 6}$$

Model M2LR is a variation of model M (eq. 1-6), except using α_R and α_P instead of α for rewards and punishments respectively in eq. 1 and 3. Model M2invT is a variation of model M except with ρ_R and ρ_P for rewards and punishments respectively in eq. 2 and 4.

In the state-independent learning model MSind, values are also allocated to actions according to the outcome of the preceding trial (eq. 7 and 8) irrespective of what cue was shown, with separate learning rate (α_{sind}) and outcome sensitivity (ρ_{sind}) parameters. Action propensities are then a weighted (via w_{sind}) combination of state-dependent ($W(t,a)$, from eq. 5) and state-independent action values (eq. 9):

$$Q_t(a_t) = Q_{t-1}(a_t) + \alpha_{sind} \times \delta_1(t) \quad \text{eq. 7}$$

$$\delta_1(t) = \rho_{sind} \times outcome(t) - Q_{t-1}(a_t) \quad \text{eq. 8}$$

$$for\ a = Go, NoGo:$$
$$W_{Msind} = (1 - w_{sind}) \times W(t,a) + w_{sind} \times Q_t(a) \quad \text{eq. 9}$$

Using hierarchical fitting and model comparison as implemented by CBM, we estimated group-level parameters for each cluster, as shown in Fig. 2C: 70 participants (8.5%) were assigned to model M; 149 (18.2%) to model Msind; 407 (49.8%) to model M2LR; and 191 (23.3%) to model M2invT.

Turning to the trial-by-trial data, Fig. 2D shows the empirical learning trajectory for each state for each of the four clusters. The probability of making a Go response should ideally increase over trials in Go+ and Go- states (solid lines should move upward) and decrease over trials in NoGo+ and NoGo- states (dashed lines should move downward). However, these curves evolve in qualitatively different manners across groups. In the Model M cluster (n = 70), participants seem only to have learnt the valence of each state, acting according to Pavlovian tendencies, with a strong predisposition to Go in rewarding states and NoGo in punishing states. This effect is also apparent in the high mean value of the inferred population-level for parameter π (Fig. 2C). Their learning trajectories (Fig. 2D, M) increase in both rewarding states (more Go in Go+ and NoGo+) and decreases in punishing states (more NoGo in Go- and NoGo) regardless of required action leading to lower performance in Go- and NoGo+ states (Fig. 2F, M). They respond more quickly with Go responses in both rewarding states compared to punishing states, providing further evidence of their high sensitivity to state valence (Fig. 2G, M). In the model M2LR cluster (n = 407), which was the largest cluster, participants learned the task in a more instrumental manner. They gradually learned to Go in Go states and NoGo in NoGo states (Fig. 2D, M2LR) and made faster Go responses in Go states and slower Go responses in NoGo states (Fig. 2G, M2LR). However, they learned unequally from rewards and punishments, with the fitted learning rate for punishing states being higher than for rewarding states ($\alpha_P > \alpha_R$; Fig. 2C). This asymmetry is reflected in a large sharp speedy decrease in P(Go) in the early trials within the NoGo- state compared to the NoGo+ state, which indicates that the value of making a Go response decreased more after a punishment compared to the omission of a reward (Fig. 2D, M2LR). In the model M2invT cluster (n = 191 participants), participants differed in terms of the variability of their choices for rewarding versus punishing states, with the fitted variability being higher for punishing than rewarding states ($\rho_P < \rho_R$; Fig. 2C). As a result, the learning trajectories for this cluster of participants flattens early for Go- and NoGo- states (Fig. 2D, M2invT) which stem from making less deterministic Go and NoGo responses in Go- and NoGo- states respectively (Fig. 2D, M2invT). This results in lower-than-anticipated performance in NoGo- across this population, which is typically considered easy to learn due to the alignment of motivation and action.

Finally, we found that over 18% of participants belonged to the model Msind cluster, adopting a strategy influenced by state-independent action learning, potentially alongside some state-dependent learning (Fig. 2A). The learning trajectories of the state-independent learners (Fig. 2D, Msind) were substantially different from those of the state-dependent learners (Fig. 2D, M/M2LR/M2invT). The probabilities of Go responses (P(Go)) in the four states displayed relatively flat curves that did not converge towards higher or lower asymptotes for Go probabilities in Go and NoGo states, respectively. Instead, they remained closely clustered together and near chance level, indicative of high response switching. In contrast to M2LR and M2invT whose longest reaction times were for incorrect Go responses in NoGo states, reaction times in the four states were consistently high and closely similar, even in Go states, suggesting the potential adoption of an equal strategy across all four states (Fig. 2G, Msind). Any minor differences observed may stem from concurrent state-dependent learning processes. In general, the reaction times of state-independent learners (were high and closely similar, even in Go states, suggesting the potential adoption of an equal strategy across all four states (Fig. 2G, Msind). Overall, the performance was lower than the two clusters who appeared to have fully learnt the 2 x 2 nature of the task (M2LR and M2invT), and even lower than cluster M, who at least were driven by the valence of the states.

Building on the findings of Ben-Artzi et al., (2022), who reported a negative correlation between state-independent learning and working memory capacity, we examined IQ scores within each cluster as a proxy for working memory capacity (Verguts & De Boeck, 2002). Calculating the median IQ scores for each cluster of participants obtained from the Wechsler Abbreviated Scale of Intelligence (WASI) revealed a difference between these groups, with participants assigned to group M and Msind exhibiting lower IQ scores (Fig. 2B). This finding aligns with their suboptimal strategy and decreased performance. Conversely, M2LR participants, who were found have the highest performance, had significantly higher IQ scores than any of the clusters (M2LR: M ($p<0.01$); M2invT ($p<0.04$); Msind ($p<0.01$)). Statistical testing was performed using non-parametric ANOVA Kruskal-Wallis test ($p<0.01$) and corrected for multiple comparisons using the Bonferroni method.

To gain greater purchase on the choices of those individuals apparently showing state-independent learning (Fig. 2D, Msind, fitted with $r_{kn} \geq 0.7$ by Msind), compared to those with state-dependent learning (Fig. 2D, M2LR and M2invT, fitted with $r_{kn} \geq 0.7$ by M2LR and M2invT), we also conducted model-agnostic analyses to identify single-trial fingerprints. On a given trial, the probability of making a Go/NoGo response should be influenced by the outcome received the last time the current cue was presented and not by previous outcomes regardless of which cue had been shown. However, for state-independent learners, responses are not solely guided by state-specific experiences; recently received outcomes biases their responses regardless of which cues were and are being shown. To examine this effect more closely, we used the fitted parameters for model Msind to compute action values based on recent past trials independent of the state, denoted as $Q_t(a)$ for each participant (eq. 7). In this way, we could identify the subset of trials where state-independent learning had a greater influence, indicated by $Q_t(a)$ being in its lower or upper tercile. To fully isolate the impact of state-independent learning, we also excluded two types of trials: 1) we omitted trials when the current and the previous state were there same; and 2) we excluded trials where the action-outcome experience on the previous trial, which now always involved a different state, was the same as

the action-outcome experience the last time the current state was visited. The upper heatmaps in Fig. 3A, B show the average probability of repeating a Go/NoGo response following a reward, neutral outcome, or punishment on the previous trial, which in this analysis always involved a different state than the current trial. By contrast, the lower heatmaps in Fig. 3C, D show the average probability of repeating a Go/NoGo response following a reward, neutral outcome, or punishment the last time the state was visited, which in this analysis is never the previous trial. The illustration shows that state-dependent learners (M2LR and M2inv) tended to repeat a rewarded action and avoid a punished action in the next visit to a specific state, and were not specifically influenced by the previous trial (compare Fig. 3B to Fig. 3D). By contrast, the state-independent learners (MSind) tended to follow the outcome associated with the last trial as opposed to the last time the current state was visited (compare Fig. 3A to Fig. 3C). The decreasing trend in probability of repeating a Go/NoGo along the x-axis in Fig. 3A shows that they learned to rely on previous trial action-outcome experience to repeat or avoid a response. This relates to their lower performance and higher decision variability. Fig 3C shows that they had nevertheless learned, to some extent, that actions and states are interconnected (Fig. 2F), suggesting parallel state-dependent (Fig. 3A) and state-independent (Fig. 3C) learning across the Msind population.

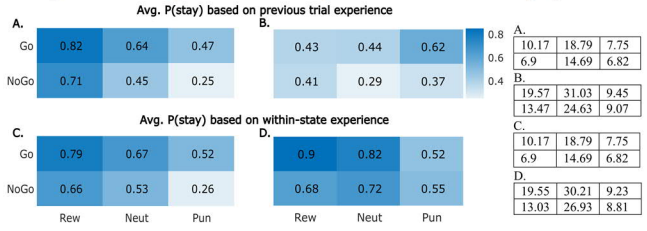

Fig. 3. Model agnostic analysis of (A and C) state-independent learners (86 participants of 149 state-independent learners with a conservative threshold of $r_{kn} \geq 0.7$) and (B and D) state-dependent learners (354 participants from M2LR andM2invT clusters with a conservative threshold $r_{kn} \geq 0.7$). In the upper rows (A-B), each cell shows the probability of repeating the same action after receiving a reward, neutral or punishment for that action in the previous trial. In the lower rows (C and D), each cell shows the probability of repeating the same action based on the performed action and received outcome in the last visit to the current state. The subset of trials used for this analysis are reported in the adjacent table as the average number of trials in each cell across participants.

To formally test these differences, we conducted a permutation test. We obtained the subset of trials for this analysis as outlined earlier and extracted for each trial, its preceding trial and last state-specific action-outcome experiences. This facilitates computing probability of repeating a Go/NoGo under the influence of outcome type in previous trial or last visit of each state. Then, we generated 100 permutations of previous trial experiences (performed actions and received outcomes in the previous trial), maintaining the integrity of within-state experiences (performed actions and received outcomes in the last visit of each state). We calculated the probability of repeating an action followed by different outcomes and compared two observed trends between the original and permuted data: 1) $P(\text{Action}|\text{Action-Rew}_{t-1}) > P(\text{Action}|\text{Action-Pun}_{t-1})$; and 2) the dominance of state-independent learning over state-dependent learning in NoGo rows meaning a tendency to repeat a NoGo action by relying more on previous trial experience than previous within-state experience. Both trends were greater in original data than in the permuted data, implying $p<0.01$, rejecting the null hypothesis that the differences in Fig. 3 were due to chance.

Finally, looking specifically at the Msind cluster, we examined how the inclusion or exclusion of the state-independent learning strategy affects the fitted parameters that are supposed to represent the behavior of this population. Although the state-independence only explains part of the behavior even for the state-independent learners, when the state-independent learning strategy is not considered (fitting Msind participants by model M), the inferences made about the parameters associated with residual state-dependent learning change substantially; such as α, π and ρ (Fig. 4), aligning more closely with the observed behavior across this population. For example, the empirical P(Go) for state-independent learners (Msind cluster) appears flat, suggesting that the estimated state-dependent learning rate (α) in model M for this group is close to zero. However, considering their average performance (Fig. 2F) and the model-agnostic analysis (Fig. 3A, C), we know that they have learned the task in a state-dependent manner to some extent, indicating that α should not be that close to zero. The inclusion of state-independent learning corrects the estimated α to better represent their partially state-dependent learning as well.

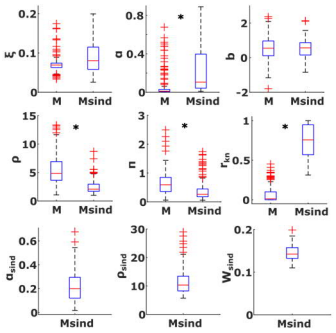

Fig. 4 Parameter estimates and responsibility ratio, as a measure of goodness of the fit in CBM, in 149 state-independent learners without (M) and with (Msind) the incorporation of state-independent learning in the model. The change in estimated learning rate (α), outcome sensitivity (ρ) and Pavlovian bias (π) is significant (marked by asterisks).

Incorporating state-independent learning to create the Msind algorithm helped to solve this issue. The consequence is that the state-dependent learning rate *LR* is estimated to be higher (making the synthetic behavior more faithful), but then the state-action values are scaled by a trade-off parameter $(1 - w_{sind})$ which are estimated based on each participant's reliance on state-dependent versus state-dependent learning. The comparison of estimated parameters between M and

Msind indicates that there is a higher degree of decision variability in state-dependent learning (as evidenced by reduced sensitivity in Msind compared to M) and that choices are more deterministic based on state-independent learning information ($\rho_{sind} > \rho$ in model Msind). These changes better capture the greater degree of switching, and, at the same time, performance, especially in Pavlovian-incongruent conditions (Go-, NoGo+), as evidenced by Fig. 2F.

Discussion

The exploration of individual differences is of importance for understanding natural variation in learning and choices and the different ways in which these processes may go awry in neuropsychiatric disorders. By applying computational modeling and behavioral analysis to a large dataset of participants performing the motivational Go/NoGo task, we identified clusters of participants characterized by distinct model-based and model-agnostic signatures.

It is possible that the four distinct clusters identified here may map onto particular clinical conditions, personal traits or neurobiological substrates. For example, participants in cluster M appear to approach the task in a Pavlovian-driven style, as also reflected in their reaction times (Fig. 2G), with faster Go responses in rewarding versus punishing states. It has been reported that disorders like impulsivity, especially in medicated Parkinson's disease (Eisinger et al., 2020) and traumatic stress (Ousdal et al., 2018), are associated with higher Pavlovian biases in motivated decision-making. The majority of participants was assigned to model M2LR, in line with the model that had previously been reported to win across the same population (Moutoussis et al., 2018). The higher learning rate for punishment was found to be correlated with higher levels of serotonin across healthy participants (Michely et al., 2022) a neurotransmitter associated with enhanced attention (Hensler, 2010) and negative outcome avoidance reported as a protective phenotype in healthy participants (Fox, Ridgewell & Ashwin, 2009). This population duly performed particularly well in the negative valence states (Go- and NoGo-), but also very well over all. A great number of participants was assigned to model M2invT, with two different outcome sensitivities. Prior research has indicated the presence of separate neural substrates for processing rewards and punishments, with clinical conditions and individual differences capable of modulating the activity within these networks (Kim et al., 2014; Tomer et al., 2014; Must et al., 2006). For instance, distinct reward-punishment sensitivity is reported to be correlated with ADHD, depression, eating disorders (Harrison et al., 2012; Portengen et al., 2021).

Although action perseveration and non-reinforcement-based effects can both lead to state-independent behavior, we found that these are distinct from the sort of learning parameterized by Msind which considers the recent action consequence independent of the context. This is evident in the decreasing trend in probability of repeating an action across outcome valences (reward, neutral and punishment) in Fig. 3A. Model-dependent and model agnostic analyses showed that learning could accrue to actions by themselves, even if it is state-action pairs that determine the actual outcomes. We hypothesize that this phenomenon becomes more prevalent under conditions of time pressure, low success rates, and imperfect task understanding, prompting participants toward statistically- and computationally-less demanding strategies. Shahar et al., (2021) has shown this approach is more common among patients suffering from OCD and Ben-Artzi et al., (2022) showed that it is negatively corelated with working memory capacity, a quantity that is itself related to IQ (Jensen, 1989; Verguts & De Boeck, (2002). The median IQ scores confirmed differentiation across the four clusters, with the highest scores observed for M2LR learners (who performed best overall) and lower scores for groups exhibiting suboptimal strategies and performance (M and Msind).

Shahar et al., (2019) reported in a 2-step task with two fractal images as offered options at each stage, that rewarding a fractal pattern increased the probability of its selection by an average of 21.4% when the fractals subsequently switched sides on the screen, compared to 37.3% when the fractal response mapping remained unchanged. They argued that the reward effect can be transferred between states as assigned value to the response key. Reward increased selection of the rewarded response key in the next trial by 4.13%. In our model agnostic analysis, we found that participants assigned to model Msind tend to repeat a rewarded Go/NoGo by 82%/71% on average and avoid it when it was punished by 47%/25% respectively. The neutral outcome can be interpreted in two ways: either as safety, indicating the absence of punishment, or as a lack of gain. In our study, the received neutral outcome in previous trial was observed to reduce the probability of repeating the previous trial action compared to reward while remained more favorable than punishment such that the probability of repeating the action that elicited nothing was higher than when punishment was received (repeating Go/NoGo if the previous trial outcome was neutral: 64%/45%). The described trend, wherein the outcome of previous trial is intricately tied to the performed action, provides evidence indicating a disregard for state information.

We confirmed the finding of state-independent learners in various ways, including: 1) conducting permutation tests on model-agnostic outcomes to demonstrate that the observed trend in action selection within the Msind cluster is not due to chance; and 2) simulating data using a model that does not have the state-independent learning component (e.g. M), finding that these synthetic participants were not captured by model Msind in model comparison and showed no sign of state-independent learning in model agnostic analysis as well. We showed that ignoring this underlying mechanism could lead to parameter estimates that significantly deviated from the characteristics observed in real data. As this extensive dataset is a component of a larger study encompassing multiple tasks within a battery, it will be intriguing to track the decision characteristics of the individuals within Msind and other clusters across other tasks.

Acknowledgment

Funding was from the Max Planck Society (AN, PD), the Humboldt Foundation (PD) and the Lundbeck Foundation (DB). PD is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764 and of the Else Kröner Medical Scientist Kolleg "ClinbrAIn: Artificial Intelligence for Clinical Brain Research" / IMPRS. IMPRS covered the conference travel expenses associated with this research (AN).

References

Ben-Artzi, I., Luria, R., & Shahar, N. (2022). Working memory capacity estimates moderate value learning for outcome-irrelevant features. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-21832-x

Cavanagh, J. F., Frank, M. J., & Allen, J. J. B. (2011). Social stress reactivity alters reward and punishment learning. *Social Cognitive and Affective Neuroscience*, *6*(3), 311–320. https://doi.org/10.1093/scan/nsq041

Dayan, P., Niv, Y., Seymour, B., & D. Daw, N. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*(8). https://doi.org/10.1016/j.neunet.2006.03.002

Eisinger, R. S., Scott, B. M., Le, A., Ponce, E. M. T., Lanese, J., Hundley, C., Nelson, B., Ravy, T., Lopes, J., Thompson, S., Sathish, S., O'Connell, R. L., Okun, M. S., Bowers, D., & Gunduz, A. (2020). Pavlovian bias in Parkinson's disease: an objective marker of impulsivity that modulates with deep brain stimulation. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-69760-y

Fox, E., Ridgewell, A., & Ashwin, C. (2009). Looking on the bright side: biased attention and the human serotonin transporter gene. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1663), 1747–1751. https://doi.org/10.1098/rspb.2008.1788

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, *62*(1). https://doi.org/10.1016/j.neuroimage.2012.04.024

Harrison, A., O'Brien, N., Lopez, C., & Treasure, J. (2010). Sensitivity to reward and punishment in eating disorders. In *Psychiatry Research* (Vol. 177, Issues 1–2). https://doi.org/10.1016/j.psychres.2009.06.010

Hensler, J. G. (2010). *Serotonin in Mood and Emotion* (pp. 367–378). https://doi.org/10.1016/S1569-7339(10)70090-4

Huys, Q. J. M., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. In *Nature Neuroscience* (Vol. 19, Issue 3). https://doi.org/10.1038/nn.4238

Kim, S. H., Yoon, H. S., Kim, H., & Hamann, S. (2014). Individual differences in sensitivity to reward and punishment and neural activity during reward and avoidance learning. *Social Cognitive and Affective Neuroscience*, *10*(9). https://doi.org/10.1093/scan/nsv007

Michely, J., Eldar, E., Erdman, A., Martin, I. M., & Dolan, R. J. (2022). Serotonin modulates asymmetric learning from reward and punishment in healthy human volunteers. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-03690-5

Mkrtchian, A., Roiser, J. P., & Robinson, O. J. (2017). Threat of shock and aversive inhibition: Induced anxiety modulates Pavlovian-instrumental interactions. *Journal of Experimental Psychology: General*, *146*(12). https://doi.org/10.1037/xge0000363

Moutoussis, M., Bullmore, E. T., Goodyer, I. M., Fonagy, P., Jones, P. B., Dolan, R. J., & Dayan, P. (2018). Change, stability, and instability in the Pavlovian guidance of behaviour from adolescence to young adulthood. *PLoS Computational Biology*, *14*(12). https://doi.org/10.1371/journal.pcbi.1006679

Must, A., Szabó, Z., Bódi, N., Szász, A., Janka, Z., & Kéri, S. (2006). Sensitivity to reward and punishment and the prefrontal cortex in major depression. *Journal of Affective Disorders*, *90*(2–3). https://doi.org/10.1016/j.jad.2005.12.005

Ousdal, O. T., Huys, Q. J., Milde, A. M., Craven, A. R., Ersland, L., Endestad, T., Melinder, A., Hugdahl, K., & Dolan, R. J. (2018). The impact of traumatic stress on Pavlovian biases. *Psychological Medicine*, *48*(2). https://doi.org/10.1017/S003329171700174X

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLoS Computational Biology*, *15*(6). https://doi.org/10.1371/journal.pcbi.1007043

Portengen, C. M., Sprooten, E., Zwiers, M. P., Hoekstra, P. J., Dietrich, A., Holz, N. E., Aggensteiner, P. M., Banaschewski, T., Schulze, U. M. E., Saam, M. C., Craig, M. C., Sethi, A., Santosh, P., Ouriaghli, I. S., Castro-Fornieles, J., Rosa, M., Arango, C., Penzol, M. J., Werhahn, J. E., … Naaijen, J. (2021). Reward and Punishment Sensitivity are Associated with Cross-disorder Traits. *Psychiatry Research*, *298*. https://doi.org/10.1016/j.psychres.2021.113795

Proulx, C. D., Hikosaka, O., & Malinow, R. (2014). Reward processing by the lateral habenula in normal and depressive behaviors. In *Nature Neuroscience* (Vol. 17, Issue 9). https://doi.org/10.1038/nn.3779

Shahar, N., Hauser, T. U., Moran, R., Moutoussis, M., Bullmore, E., Dolan, R. J., Goodyer, I., Fonagy, P., Jones, P., Moutoussis, M., Hauser, T., Neufeld, S., Romero-Garcia, R., Clair, M. S., Vértes, P., Whitaker, K., Inkster, B., Prabhu, G., Ooi, C., … Dolan, R. J. (2021). Assigning the right credit to the wrong action: compulsivity in the general population is associated with augmented outcome-irrelevant value-based learning. *Translational Psychiatry*, *11*(1). https://doi.org/10.1038/s41398-021-01642-x

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., Nspn, C., & Dolan, R. J. (2019). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32). https://doi.org/10.1073/pnas.1821647116

Tomer, R., Slagter, H. A., Christian, B. T., Fox, A. S., King, C. R., Murali, D., Gluck, M. A., & Davidson, R. J. (2014). Love to win or hate to lose? Asymmetry of dopamine D2 receptor binding predicts sensitivity to reward versus punishment. *Journal of Cognitive Neuroscience*, *26*(5). https://doi.org/10.1162/jocn_a_00544

Verguts, T., & De Boeck, P. (2002). On the correlation between working memory capacity and performance on intelligence tests. *Learning and Individual Differences*, *13*(1), 37–55. https://doi.org/10.1016/S1041-6080(02)00049-3