Model-based prioritization for acquiring protection

Sarah M. Tashjian[a,*], Toby Wise[a,b], & Dean Mobbs[a,c]

[a]Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, USA
[b]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[c]Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA

*Corresponding author Sarah M. Tashjian (smtashji@caltech.edu), Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, MC 228-77, Pasadena, CA 91125, USA

Author Contributions
SMT developed the study concept with input from DM. SMT designed the study with input from DM and TW. Data collection was performed by SMT. Data analysis and interpretation were performed by SMT and TW. SMT drafted the manuscript, with critical revisions from DM. All authors approved of the manuscript.

Competing Interests
The authors declare no competing interests.

Keywords: model-based control, protection, punishment, reinforcement learning, safety

This pdf file includes:
Main text
Figures 1 to 4
Supplemental Materials

Abstract

Protection, or the mitigation of harm, often involves the capacity to prospectively plan the actions needed to combat a threat. The computational architecture of decisions involving protection remains unclear, as well as whether these decisions differ from other beneficial prospective actions. Here we compare protection acquisition to reward acquisition and punishment avoidance to examine overlapping and distinct features across the three action types. Protection acquisition is positively valenced similar to reward acquisition. For both protection and reward, the more the actor gains, the more benefit. However, reward and protection occur in different contexts, with protection existing in aversive contexts. Punishment avoidance also occurs in aversive contexts, but differs from protection because punishment is negatively valenced. Across three independent studies (Total *N*=600) we applied computational modeling to examine model-based reinforcement learning for protection, reward, and punishment in humans. Decisions motivated by acquiring safety via protection evoked a higher degree of model-based control than acquiring reward or avoiding punishment, with no significant differences in learning rate. The context-valence asymmetry characteristic of protection increased deployment of flexible decision strategies, suggesting model-based control depends on the context in which outcomes are encountered as well as the valence of the outcome.

Significance Statement

Acquiring protection is a ubiquitous way humans achieve safety through prospective decisions. Prospective safety decisions likely engage model-based control systems, but computational decision frameworks have yet to be applied to safety. Clinical science instead dominates investigations of safety by examining protection acquisition in its maladaptive form. The current studies apply computational models of decision control systems identified in reward learning to understand how humans make adaptive decisions to acquire protection. Distinctive context and valence qualities of protection were reflected in increased model-based control for safety-motivated decisions compared to reward- or threat-motivated decisions, despite superficial similarities. Inability to effectively use model-based control may reveal new insights into how safety decisions go awry in psychopathology.

Humans have a remarkable capacity to foresee and avoid harm through protective strategies.[1] Acquiring protection through prospective decisions is a predominant way humans achieve safety: we build fences to keep out dangerous animals, buy weapons to defend against conspecifics, and wear protective clothing to shield from natural elements. Prospective actions like these protective behaviors likely utilize model-based decision control systems, which support goal-directed action. The relative contributions of decision control systems to behavior remains an ongoing topic of research. To our knowledge no prior work has examined how computational control systems support adaptive protection acquisition. Extant literature on decision making focuses on purely appetitive or aversive domains, or, when considering safety, focuses on maladaptive safety decisions. These approaches fail to precisely consider adaptive safety decisions, which are ubiquitous and important for human functioning. We address this gap by matching protection acquisition with reward acquisition and punishment avoidance decisions using a novel set of tasks designed to engage model-based and model-free control systems.

Computational reinforcement learning frameworks are powerful methods for characterizing decision making that have yet to be applied to protection. Existing literature provides strong evidence that reward- and punishment-motivated decisions are underpinned by model-based and model-free control systems.[2,3] With model-free control, actions that are positively reinforced are repeated when similar stimuli are subsequently presented. The resulting habit-like actions are stimulus-triggered responses based on accumulated trial and error learning rather than prospective, goal-directed responses. The model-based system, in contrast, builds a map of the environment and uses that map to determine the best course of action based on environmental structure. The model-free system is computationally efficient whereas the model-based system is computationally intensive but highly flexible.[4] Individuals tend to use a mix of strategies with differences dependent on task demands,[5] development,[6] and psychiatric symptomology.[7]

Protective decisions retain superficial similarities to both reward- and punishment-motivated decisions. Protection is positively-valenced similar to reward but exists in a negative context similar to punishment (Fig. 1a). Motivationally, negatively-valenced stimuli like punishment typically elicit avoidance behaviors, whereas positively-valenced stimuli like reward elicit approach behaviors.[8] Unlike previous studies of learning in aversive contexts,[9,10] protection in this study did not center on avoidance. Instead, subjects were incentivized to actively seek out the maximum protection available as opposed to avoid the highest punishment. This aligns with traditional definitions of approach motivation as the energization of behavior toward a positive stimulus.[11] However, protection differs from purely appetitive or aversive stimuli that are matched on valence and context. The degree to which valence and behavior are aligned (i.e., approach for reward versus avoid for reward) has consequences for learning,[12] but prior studies of decision control typically exploit the conventional coupling of valence and context. Applying computational models of decision control in purely appetitive or aversive domains does not sufficiently identify how decision control systems contribute to acquiring protection.

It remains an open question whether shared valence with reward or shared context with punishment results in a shared computational decision structure for protection. In prior work, reward acquisition and punishment avoidance elicit similar weighting of model-based control.[9,13] Reward and punishment share overlapping valence with the context in which they are situated. This valence-context symmetry[14] should favor model-free control as a result of less complex contingency learning.[5,12] In predictable environments, reward learning engages goal-directed control early on, but cedes to habitual control as an efficiency.[15] Prospective model-based control facilitates the development of accurate and flexible action policies,[9] which may be more necessary when valence and context disagree, as is the case with protection. Thus, the valence-context

asymmetry of protection may bias toward greater prospective model-based control than both comparison stimuli.

Identifying the extent to which model-based control differs for protection acquisition compared with reward and punishment is an important step in understanding how humans make adaptive safety-motivated decisions. Effectively mapping protection contingencies through model-based control is posited to create a positive feedback loop: prospective safety extends the capacity for evaluating and integrating knowledge about the environment when later circumstances limit time to engage the model-based system.[16] Over-reliance on model-free systems, by contrast, may underlie pathological as opposed to adaptive protection-seeking, which is characterized by repeating safety behaviors that are disproportionate to the threat faced.[17] An unresolved question centers on whether valence and context features of protection relate to the degree to which an individual's actions reflect the dominance of the model-based system or the more reflexive and habitual model-free control system.

In the current set of three preregistered studies (*N*=600 total), we applied computational modeling to characterize learning for positively-valenced stimuli in disparate contexts (Study 1 and 2) and learning for disparately-valenced stimuli in the same context (Study 3). We developed five modified versions of a widely-used two-step reinforcement learning task examining protection acquisition (positive valence, aversive context), reward acquisition (positive valence, positive context), and punishment avoidance (negative valence, aversive context).[18] We hypothesized that the context-valence asymmetry for protection would increase model-based contributions compared to reward (Study 1 and 2) in line with prior work showing aversive contexts decrease model-free contributions to reward learning.[19] We hypothesized that model-based contributions for protection would also be higher compared to punishment avoidance given the potential for combined contributions of appetitive and aversive motivations for protection (Study 3). We examined effects of incentives to determine whether differences in model-based control were modulated by incentive (i.e., are protection and reward only differentiated by the model-based system when stakes are high).[18,20] We hypothesized that incentive sensitivity would be higher for reward given lower value thresholds for protection and punishment. Lastly, we examined metacognitive and predictive accuracy on each task to determine how metacognitive awareness of task performance related to model-based control. We hypothesized increased model-based control would be associated with better metacognitive accuracy across all stimuli.

Results

In each pre-registered study, a balance between model-free and model-based control was assessed using two variants of a two-step reinforcement learning task. Each study included a protection task variant and either a reward or punishment task comparison. During each task, subjects made sequential decisions that navigated them through two "stages" defined by different stimuli. Subjects were told they were traveling through a fictious forest. At the first stage they had to choose a dwelling to visit. Each dwelling led to a second-stage creature that was associated with a probability of receiving an outcome. At the first-stage, two equivalent states were randomly presented throughout and each first-stage state included two dwellings (4 total dwellings). In each of the first-stage states, one dwelling led to one creature and the second dwelling led to a different creature (2 total creatures), creating an implicit equivalence across first-stage states. In the protection task variant the second-stage outcomes were protection stimuli that reduced losses. In the reward task variants the second-stage outcomes were reward stimuli that increased gains. In the punishment task variant the second-stage outcomes were punishment stimuli that increased losses. Second-stage probabilities changed slowly over time, requiring continuous learning in order select the appropriate first-stage state that led to optimized outcomes.

To quantify the computational mechanisms underpinning behavior, we fit four computational models to subjects' choice data during the task (see "Methods" for a full description of all tested models). The parameter of primary interest was the balance between model-free and model-based control between task variants. At an individual level, this balance can be quantified by a hybrid model, which combines the decision values of two algorithms according to a weighting factor (ω). A learning rate (α) parameter was also estimated for integrating outcomes to update choice behavior. Additional model-parameters included a single eligibility trace (λ), stickiness (π) and inverse-temperature (β) parameter. Watanabe-Akaike Information Criterion (WAIC) scores were used as a complexity-sensitive index of model fit to determine the best model for each study.

To validate the computational modeling analyses, we used mixed-effects logistic regressions to examine choice behavior as a function of the outcome on the previous trial and similarity in first-stage state. Choice behavior was measured as the probability of repeating a visit to the same second-stage state ("stay probability"). The interaction between first-stage state and previous outcome indicates a model-free component whereas a main effect of previous outcome indicates a model-based component.[21] For the model-free strategy, outcomes received following one first-stage state should not affect subsequent choices from a different first-stage state because an explicit task structure is not mapped (thus the equivalence between first-stage states is not learned). The model-free learner only shows increased stay probability when the current first-stage state is the same as the first-stage state from the previous trial, and this is reflected as an interaction between previous outcome and first-stage state. The model-based learner, in contrast, uses the task structure to plan towards the second-stage outcomes, allowing it to generalize knowledge learned from both first-stage states. Thus, outcomes at the second stage equally affect first-stage preferences, regardless of whether the current trial starts with the same first-stage state as the previous trial.

Study 1: Comparing protection acquisition to reward acquisition. In Study 1, two-hundred subjects (M_{age}=27.99(6.87), $range_{age}$=18-40 years, 98 females, 102 males) completed the protection acquisition and reward acquisition task variants (Fig. 1b and c). The aim in both variants was to earn the maximum possible second-stage outcome thereby optimizing the final result. Second-stage outcomes for the protection acquisition variant were shields that served as protective stimuli to reduce punishment in the form of dragon flames. Second-stage outcomes for the reward acquisition variant were sacks that served as reward stimuli to increase reward in the form of fairy coins (Fig. 1c). The number of second-stage outcomes earned ultimately affected the final result, which was points that contributed to subjects' bonus payments.

Task Engagement. Subjects were engaged and performed well, as shown by higher than median available outcomes earned (Fig. S1a). Average number of points earned per trial (reward rate) was calculated for each subject and mean corrected by the average outcome available to them via individually generated point distributions. Corrected reward rate was higher on the protection task variant and did not differ as a function of stakes within each task variant (Fig. S1b). Subjects made first-stage decisions in less than 1 second, and first-stage reaction time (RT) was significantly faster for the protection variant (Fig. S1c).

Computational Models. The best fitting model was Model 3, which included separate model-based weighting (ω) and learning rate (α) parameters for each task variant, as well as eligibility trace (λ), stickiness (π) and inverse-temperature (β) parameters (Table S1). Subjects had higher average ω on the protection variant compared to the reward variant (Fig. 2a). α did not significantly differ by task variant (Fig. 2b).

To confirm a higher degree of model-based decision making led to better performance, we examined associations between corrected reward rate and ω. Higher ω-value led to better performance (Fig. 2c). Higher α was also associated with better performance (Fig. 2d). No moderation by task variant was observed in analyses controlling for task order and variant.

Mixed-Effects Models. A significant main effect of previous outcome on staying behavior was observed, indicating contributions from model-based control (Fig. 2e). An interaction between first-stage state and previous outcome was also significant, indicating contributions from a model-free component. Lastly, a three-way interaction with task variant was observed confirming reinforcement learning results that subjects were more model-based on the protection task variant. In other words, prior outcome had a stronger effect on stay behavior for the protection variant compared to the reward variant when the first-stage states differed from the previous trial, but not when the first-stage state was the same (Fig. 2f).

Reinforcement learning models did not reveal a meaningful stakes effect (Model 4 fit was not significantly better than Model 3). Stakes also did not moderate stay behavior with either the model-based or model-free mixed-effects components (Fig. S2a). These results were consistent for each task variant individually. Because this null effect of stakes differed from prior work using a similar task and our task included fewer trials compared with prior work,[20] we explored whether the effect of stakes varied as a function of task duration. Task duration interacted with stakes and previous outcome, such that there was no effect of stakes at the start of the task but high-stakes trials had an increase in likelihood of stay behavior at the end of the task (Fig. S2b).

Study 2: Comparing protection acquisition to direct reward. In Study 2, two-hundred subjects (M_{age}=23.26(4.02), $range_{age}$=18-36 years, 152 females, 48 males) completed the protection acquisition and direct reward acquisition task variants (Fig. 1b and d). Study 2 was conducted with the same protection acquisition task as Study 1, but substituted the reward acquisition task with a modified version to test whether indirect reward delivery in Study 1 (i.e., learning to acquire sacks rather than coins themselves) artificially reduced model-based control contributions to reward acquisition (Fig. 1d). Study 2 also increased the number of trials for both task variants. Study 1 did not reveal an effect of stakes, as found in prior work examining reward learning,[20] but Study 1 had a fewer number of trials than prior work. We thus tested whether the lack of stakes effect in Study 1 was due to the number of trials by increasing the number of non-practice trials in Study 2 from 100 to 200 for each variant, consistent with prior work.[20] Again, the aim in both task variants was to earn the maximum possible outcome (shields, coins) thereby optimizing the final result (points contributing to bonus).

Task Engagement. As in Study 1, subjects were engaged and performed well, with higher than median outcomes earned, RTs of <1 second, and higher corrected reward rate for protection compared to direct reward (Fig. S1a-c). No stakes effect was observed for corrected reward rate.

Computational Models. Computational Model 4 was the best fitting model for Study 2, which included separate ω and α parameters for task variant and stakes. ω was higher for the protection variant compared to the direct reward variant, consistent with Study 1 (Fig. 2a). Diverging from Study 1, ω differed between tasks for both high- and low-stakes trials, with the protection variant demonstrating more model-based control for both stakes (Fig. S3). As in Study 1, α did not significantly differ by task variant (Fig. 2b). See Table S1 for λ, π, and β parameters. Consistent with Study 1, higher ω and α parameters were associated with better performance as indexed by increased corrected reward rate, with no moderation by task variant (Fig. 2c-d).

Mixed-Effects Models. Mixed-effects models identified model-based and model-free contributions (Fig. 2e), and replicated the moderation by task variant observed in Study 1 (Fig. 2g). Stakes interacted with the model-based component, which was driven by the direct reward variant (Fig. S2c). No significant stakes interaction was present for the model-free component.

Study 3: Comparing protection acquisition to punishment avoidance. In Study 3, two-hundred subjects (M_{age}=22.25(3.85), $range_{age}$=18-39 years, 159 females, 41 males) completed the protection acquisition and punishment avoidance task variants (Fig. 1b and e). The longer protection acquisition variant from Study 2 was used. The punishment avoidance variant

was a traditional aversive avoidance variant where the aim was to avoid punishment (dragon flames) that was delivered directly at stage-two (Fig. 1e).

Task Engagement. Outcomes earned were higher than median available outcomes, RTs were <1 second, and subjects earned more for protection compared to punishment avoidance (Fig. S1a-c). No stakes effect was observed for corrected reward rate. RT for first-stage decisions differed by stakes, such that RTs were slower for high stakes trials (Fig. S4).

Computational Models. As in Study 1, the best fitting model was Model 3, which included separate ω and α parameters for task variant, but not for stakes type. Subjects had higher ω parameters on the protection variant compared to the punishment avoidance variant (Fig. 2a). α was significantly higher for protection compared to punishment avoidance, revealing the only learning-rate difference by task variant across the studies (Fig. 2b). See Table S1 for λ, π, and β parameters. Consistent with Studies 1 and 2, higher ω and α led to better performance, as indexed by corrected reward rate, with no moderation by task variant (Fig. 2c-d).

Mixed-Effects Models. Mixed-effects models identified model-based and model-free contributions (Fig. 2e), and replicated the moderation by task variant observed in Studies 1 and 2 (Fig. 2h). No stakes effect was observed with respect to either model-based or model-free component (Fig. S2d).

Computational Model Parameter Comparisons Between Studies. Model-based control (ω) for the protection variant did not significantly differ across studies, $F(2, 597)=.95$, $p=.386$. For non-protection task variants (i.e., reward and punishment avoidance), ω differed across studies, $F(2, 597)=21.95$, $p<.001$. Post-hoc Tukey HSD comparisons revealed a significant difference between reward in Study 1 and direct reward in Study 2 as well as punishment avoidance in Study 3, such that both Study 2 and 3 ω were .13 higher than Study 1, $p<.001$. Direct reward in Study 2 and punishment avoidance in Study 3 did not differ, $w_{diff}=-.006$, $p=.963$.

Learning rate (α) significantly differed across studies for the protection task variants, $F(2, 597)=6.48$, $p=.002$, as well as non-protection task variants, $F(2, 597)=5.12$, $p=.006$. Post-hoc Tukey HSD comparisons revealed a significant difference between protection in Study 3 and protection in Study 1 as well as protection in Study 2, such that Study 3 α was .12 higher than Study 1, $p=.001$, and .08 higher than Study 2, $p=.038$. Study 1 reward α was .10 lower than Study 2 direct reward, $p=.008$, and .09 lower than Study 3 punishment, $p=.036$.

Metacognitive and Predictive Accuracy. Model-based actions can be implicit, where there is a nonconscious anticipation of an outcome, or explicit, where a conscious prospection can motivate behavior.[22] We assessed measures of metacognitive accuracy (certainty) and predictive accuracy (outcome estimates) to examine whether explicit evaluation tracked outcomes and whether accuracy differed as a function of learning strategy and context.

Average certainty and outcome estimate ratings on a scale of 0-9 were above the midpoint for all task variants (Fig. 3a-b). Mixed effects regression results indicated that certainty and outcome estimates were higher on trials for which subjects earned higher outcomes, reflecting metacognitive and predictive accuracy, respectively (Fig. 3c).

Random slope coefficients were extracted from the model of outcome predicting certainty and outcome estimates. Coefficients represented metacognitive and predictive accuracy for each subject. Metacognitive accuracy only differed by task variant for Study 3, with higher accuracy for protection than punishment avoidance (Fig. 3d). Predictive accuracy was higher for protection than reward and punishment in Studies 1 and 3, but higher for direct reward compared to protection in Study 2 (Fig. 3e). Accuracy coefficients were regressed against ω and α for each task variant. Higher metacognitive accuracy was associated with faster learning rate for all tasks and more model-based control for protection and reward in Studies 1 and 2 (Fig. 3f). Predictive accuracy was inconsistently related to model-based control, but was related to faster learning rate across all task variants with the exception of protection in Study 1 (Fig. 3f).

Anxiety. All subjects provided self-report assessments of anxiety using the State-Trait Anxiety Inventory (STAI) Trait Anxiety subscale.[23] Average scores were *M*=29.59, *SD*=11.37, range=1-58. Differences in deployment of model-based control were associated with anxiety such that individuals with higher scores on the STAI demonstrated greater model-based weighting for reward acquisition compared with protection acquisition, but greater model-based weighting for protection acquisition compared with punishment avoidance: study by ω interaction Estimate=5.40, *SE*=2.21, *t*=2.45, *p*=.015, 95% CI [1.07, 9.74], R^2=.02 (Fig. 4). Anxiety was not significantly associated with learning rate differences: Estimate=1.38, *SE*=1.66, *t*=.83, *p*=.406, 95% CI [-1.88, 4.63], R^2=.01.

Anxiety interacted with task type such that individuals with higher anxiety reported reduced certainty for protection acquisition compared to reward and direct reward, but increased certainty for protection acquisition compared to punishment avoidance (Fig. S5a). Subjects with higher anxiety also demonstrated lower outcome estimates for protection acquisition compared with direct reward, but higher outcome estimates for protection acquisition compared to punishment avoidance (Fig. S5b). Anxiety was not significantly associated with metacognitive or predictive accuracy.

Discussion

This is the first study we are aware of that determines how humans apply reinforcement learning strategies to adaptively acquire protection. Reinforcement learning models describe how predictions about the environment facilitate adaptive decision making. In aversive contexts, predictions center on minimizing harm whereas appetitive contexts motivate reward maximization. Traditional safety conceptualizations center on threat and, as such, typically elicit avoidance as opposed to approach behavior.[24] However, the current study required approach behavior to maximize positively-valenced protection. Our results demonstrate that individuals engaged model-based control systems to a greater extent when acquiring protection compared to acquiring reward and avoiding punishment. In contrast, learning rates did not differ between contexts or by stimulus valence and represented a stable individual difference. By reconceptualizing safety in terms of appetitive protection, this study progresses understanding of context-valence interactions underlying differential recruitment of decision control systems.

Situating safety in an approach orientation by using protection revealed the existence of increased flexible goal pursuit based on asynchronous context and stimulus valence. Approach motivation is frequently defined as the energization of behavior toward a positive stimulus.[11] Although the ultimate goal of protection acquisition is typically oriented toward harm avoidance, in contrast to reward acquisition which does not explicitly consider harm, protection and reward in this study were both positively valenced (i.e., more is better). Consequently, our paradigm more closely equates protection with representation of reward outcome states than prior studies with traditional aversive stimuli, as well as punishment avoidance in Study 3. Notably, greater model-based control was yoked to more beneficial outcomes through deterministic state transitions.[21] This design feature clarifies prior work that suggested greater reliance on model-based control may be suboptimal in aversive contexts.[9] When the goal is to maximize protection, and deployment of model-based control can better achieve that goal, individuals show increased reliance on model-based control.

Using both computational modelling and model-agnostic analyses, our findings revealed that protection amplifies contributions from the model-based system when compared with traditional appetitive reward and aversive punishment. Thus, aversive context can motivate flexible and adaptive behavior for positively-valenced outcomes. Differences in model-based control were not attributable to task complexity as evinced by faster reaction time for protection than other task variants and no significant differences in learning rate between tasks. Thus, it is

not that protection as a construct is more abstract thereby requiring more effortful control to optimize behavior. Instead, these results suggest value-based features of protection drive engagement of more flexible, goal-directed learning systems.

Metacognitive and predictive accuracy were higher for acquiring protection than avoiding punishment. In line with the reinforcement learning results, it is likely that punishment avoidance engages reflexive decision circuits, whereas protection acquisition more closely approximates distal threat affording engagement of cognitive circuits.[16] These findings clarify prior work identifying a reduction in metacognitive accuracy in aversive contexts,[25] and provide support for our assertion that approach and avoidance motivations underlie differences previously attributed to context. Because our design improved accuracy-demand trade-offs through deterministic state transitions, our results are also consistent with recent work showing higher decision confidence is associated with model-free learning when model-based and model-free systems have chance level performances.[26] In this study, better metacognitive accuracy was associated with more model-based control and faster learning rates. Thus, model-based actions can be interpreted to reflect explicit forecasts. Predictive accuracy demonstrated the same general pattern with less consistency, perhaps because of increased noise in the predictive accuracy measurement and the inherent difficulty in estimating precise outcomes with random walks. Together, metacognitive accuracy findings suggest that a boost in model-based control by reframing safety as an approach toward protection mitigates metacognitive differences previously linked to context.

Individual differences in trait anxiety were associated with degree of model-based control deployed to acquire protection, offering a potential mechanistic explanation for differences in safety decisions previously documented in anxious individuals.[27] For individuals with higher anxiety, model-based control for protection was decreased compared with reward, but elevated when compared with punishment. This increase in model-based control across the valence spectrum also supports our assertion that protection acquisition is distinct from punishment avoidance, and that safety exists on the spectrum between reward and threat. Trait anxiety was also associated with a general reduction in certainty and outcome estimates across the valence spectrum, but not with decreased metacognitive or predictive accuracy. Together, these finding suggests that individuals with higher anxiety performed worse on tasks involving negative context, but that they were able to estimate their performance with comparable accuracy to those with lower anxiety scores. The reduction in model-based control for protection has implications for real-world behaviors observed in anxiety. Model-free responses to protection acquisition can lead to repeating overly-cautious avoidance behaviors, often referred to as problematic “safety-seeking”. Model-based control, however, can facilitate flexible updating in response to changing threat contingencies, which supports adaptive safety acquisition. The increase in strategic control for protection compared with punishment raises the possibility that leveraging approach motivation may be beneficial for protection-based learning in anxious individuals. Applying computational decision frameworks to safety extends understanding of how humans rationally acquire protection in the face of threat and how decision control strategies differ compared with other appetitive and aversive stimuli.

This study has implications beyond informing theoretical frameworks to potentially expanding clinical approaches to treating anxiety. Up to 50% of individuals with anxiety do not fully respond to current treatments (e.g., cognitive behavioral therapy).[28] This may be, in part, due to clinical science conceptualizations of safety seeking as dysfunctional avoidance[29,30] contributing to the onset and maintenance of anxiety.[31] However, recent work proposes focusing on learning about safety, as opposed to threat, may be a promising alternative avenue by which to improve anxiety treatment.[32,33] The current work supports this call to disaggregate threat extinction and safety acquisition. Our findings show adaptive safety acquisition does not function the same as threat avoidance, even for those with higher trait anxiety. Similar to conditioned

inhibition approaches, our findings indicate safety via protection can be trained in the presence of threat thereby reducing competing associations formed during Pavlovian extinction learning.[32] The current paradigm also has the potential to differentiate between maladaptive coping strategies such as avoidance and adaptive coping strategies[34] such as flexible safety maximization by examining decisions to acquire protection when doing so is rational and in the presence of threat.

Study findings should be considered in the context of design limitations. The sample was recruited and tested online without a primary aversive stimulus (i.e., shock). Monetary outcomes have been previously validated for studying aversive and appetitive learning,[13,18,22] but other aversive contexts will need to be tested to enhance generalizability. Although we based our paradigm development on widely-used and validated reinforcement-learning tasks, we only replicated the stakes effect observed in prior work in Study 2.[20] The model accounting for both task variant and stakes fit best for Study 2, but the WAIC score for the more complex model was only .18% different from the simpler model. The simpler model accounting for task variant but not stakes fit best for Studies 1 and 3. Null findings for increased incentives (i.e., stakes) across other task variants raise the possibility that other task-based factors not yet classified in the reinforcement-learning literature are at play.[21] Additional tasks exploring approach-based safety are needed to further validate and replicate the constructs examined here. We did not collect race and ethnicity data for our sample, which precludes conclusions as to whether our sample adequately represents the broader population. We also did not assess intolerance of uncertainty, which is considered a lower order factor related to anxiety, and is often correlated with trait anxiety, but has independent predictive value.[35] Intolerance of uncertainty has been shown to specifically relate to physiological regulation in response to threat and safety cues during conditioning.[36,37] With regard to decision control systems, prior findings suggests model-free control may be more optimal under high uncertainty, particularly for punishment avoidance.[9] However, more work is needed to understand how dispositional intolerance of uncertainty interacts with situational uncertainty to influence decision control systems and learning.

Every day, humans seek to acquire protection through prospective decisions, which engage model-based decision control systems. The current studies illuminate computational decision control components that differentiate protection acquisition from reward acquisition and punishment avoidance. We focus on how humans make adaptive decisions to seek out protective stimuli as a rational choice behavior when threat is present. This focus on beneficial safety seeking differs from examinations of aberrant safety seeking, which currently dominates the literature given important ties to psychopathology. Here we provide evidence that the valence and context asymmetry of protection increased goal-directed control compared with other stimuli that have consistent valence and context matching (i.e., reward and punishment). By identifying the engagement of flexible decision control systems in protection acquisition, this work lays the foundation for better understanding of how humans adaptively acquire safety and how safety learning goes awry in psychopathology.

Methods

Sample. Six-hundred human subjects completed two reinforcement learning tasks online. Study 1 compared the protection and reward variants, Study 2 compared a longer version of the protection variant and a direct reward variant, and Study 3 compared the Study 2 protection variant and a punishment avoidance variant. Each study involved an independent sample of 200 subjects. Subjects were recruited through Prolific, an online recruitment and data collection platform that produces high-quality data.[38] As described in our preregistration, we used a stopping rule of 200 subjects with useable data. In each Study, subjects completed two task variants (90-minutes, Study 1; 120-minutes, Studies 2 and 3).

Inclusion and Exclusion Criteria. Subjects were included based on being aged 18-40, fluent in English, and normal or corrected vision. Subjects were excluded from all analyses and replaced through subsequent recruitment if they failed to respond to more than 20% of trials within the allotted time or if they incorrectly responded to more than 50% of comprehension checks. In total, 15 subjects (2.5% of the total sample) failed these criteria (3 Study 1, 8 Study 2, 4 Study 3) and were replaced through subsequent recruitment.

Ethics. All methodology was approved by the California Institute of Technology Internal Review Board, and all subjects consented to participation through an online consent form at the beginning of the experiment. Subjects were compensated for their time at a rate of US$9.00 per hour and were entered into a performance-contingent bonus lottery for US$100.00. The lottery served to increase task engagement.

Materials and Procedure. In Study 1, subjects played two similar games in which they were traveling through a forest with a goal to either maximize protection (protection variant) or reward (reward variant). Each variant consisted of 120 trials, with the first 20 trials designated as practice and not included in analyses. In Study 2, subjects played a longer version of the same protection variant and a modified version of the reward variant with reward directly delivered at Stage 2 (direct reward variant). In Study 3, subjects played the longer protection variant and a punishment variant with the goal to minimize punishment (punishment avoidance). In Study 2 and 3, trial numbers were increased to 225 trials, with the first 25 designated as practice in line with prior work examining the effect of stakes.[20] Presentation of the two task variants were counterbalanced across subjects for each study.

Prior to completing the tasks, subjects were instructed extensively about the transition structure, outcome distribution, and how the stakes manipulation worked. Subjects completed 10 comprehension questions (no time limit) with feedback after the task instructions. Subjects were excluded from analyses and replaced through subsequent recruitment if they completed less than 50% of comprehension questions accurately. Instructions and comprehension were included to ensure subjects fully understood task elements.

Stakes. Each trial started randomly with an indicator of high (x5) or low (x1) stakes for 1500ms. Specifically, in all protection and the punishment avoidance variants subjects were shown dragons who were either small and delivered one flame or were large and delivered five flames. In the reward and direct reward variants subjects were shown fairies that were either small and delivered one gold coin or were large and delivered five gold coins. Large stakes distributed 45 units of reward/punishment whereas small stakes distributed 9. This allowed for an outcome/reward distribution similar to that used previously with varying stakes magnitudes.[18]

First-stage choices. After the stakes depiction, one of two possible first-stage states was randomly shown. In all protection variants, first-stage states were depicted as trees where gnomes dwelled. In all other variants, first-stage states were depicted as houses where elves dwelled. Houses were used to denote reward, direct reward, and punishment in order to minimize between-study differences in comparisons as a function of stimulus kind. Subjects had to choose between the left- and right-hand first-stage dwellings using the “F” and “J” keys within a response deadline of 1500ms. If subjects did not respond within the time allotted, they were told they did not select in time and were instructed to press “space” to start the next trial.

Second-stage outcomes. First-stage choices determined which second-stage state was encountered. Deterministic transitions specified that the same first-stage dwelling always led to the same second-stage state, which was depicted as a creature. Choices between first-stage states were equivalent between such that a dwelling in each pair always led to one of the two second-stage creatures and the other always led to the remaining creature. This equivalence distinguished model-based and model-free strategies because only the model-based system can transfer learned experiences from one first-stage state to the other.[21] This is an important aspect

of the task given growing evidence that both model-free and model-based strategies can result in optimal decisions depending on task constraints.[39] In the Study 1 reward variant, second-stage creatures were elves who made sacks to carry the gold coins. In the Study 2 direct reward and Study 3 punishment avoidance variants, coins and flames were delivered at the second-stage. In all protection variants, second-stage creatures were gnomes who made shields to protect against dragon flames. Payoffs were initialized as low (0-4 points) for one creature and high (5-9 points) for the other. Payoffs changed slowly over the course of the task according to independent Gaussian random walks (σ=2) with reflecting bounds at 0 and 9 to encourage learning throughout. A new set of randomly drifting outcome distributions was generated for each subject.

Final result. After making their first-stage choice, subjects were shown which creature they visited for 1500ms and then were shown how much outcome they received as well as the final result based on how outcomes were applied to the initial stakes for 2500ms. Subjects lost points if attacked in the protection and punishment avoidance variants and gained points in the reward and direct reward variants. Points contributed to actual bonus money distributed. Similar incentives have successfully been used in prior work.[10]

Reinforcement Learning Models. Reinforcement learning models were fit using a hierarchical Bayesian approach, assuming subject-level parameters are drawn from group-level distributions, implemented in Stan,[40] which allowed us to pool data from all subjects to improve individual parameter estimates. Building on prior work,[18,21] we fit 4 computational reinforcement learning models with weighting parameters (ω) that determined the relative contribution of model-based and model-free control and learning-rate parameters (α) that governed the degree to which action values were updated after a positive outcome. Models also included an eligibility trace parameter (λ) that controlled the degree to which outcome information at the second stage transferred to the start stage, a "stickiness" parameter (π) that captured perseveration on the response, and an inverse-temperature parameter (β) which controlled the exploitation exploration trade-off between two choice options given their difference in value. Model fitting was conducted as follows: (Model 1) First we fit a "null model" that did not include an effect of stakes or task variant and accounts for subjects' choices by integrating first-stage value assignment for both model-based and model-free systems. Four distinct first-stage states were assumed. (Model 2) Next, we fit a model that included the same first-stage model-based and model-free learning as Model 1 with an additional separate ω and α parameter for the effect of high- and low-stakes trials. (Model 3) Then, we fit a model that included the same first-stage model-based and model-free learning as Model 1 with an additional separate ω and α parameter for each task variant. (Model 4) Finally, we fit a model that included the same first-stage learning and task variant effect as Model 3 with an additional separate ω and α parameter for the effect of high- and low-stakes trials. Model-fitting was performed using weakly informative prior distributions (normal distributions with mean=0 and standard deviation=1). Posterior distributions for model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling implemented in Stan, with 4 chains of 4000 samples each. For further analyses, we used the mean of each parameter's posterior distribution. Model comparison was performed using Watanabe-Akaike Information Criterion (WAIC) scores,[41] which provides a goodness of fit measure for Bayesian models penalized according to the number of free parameters in the model. Lower WAIC scores indicate better out-of-sample predictive accuracy of the candidate model.

Mixed-Effects Models. All statistical tests, with the exception of the reinforcement learning models, were conducted in R (version 4.0.3) using the lme4 package (version 1.1.26).[42] Mixed effects models were tested using the lmer function (lmerTest assessed t-tests using Satterthwaite's method). Linear models were tested using the lm function. General effects sizes are reported as 95% confidence intervals. Model effect sizes reported as R^2 are conditional effects of variance explained by the entire model.[43]

Metacognitive and Predictive Accuracy. Decision certainty and outcome estimates were collected throughout the tasks on 25% of trials each. Subjects were asked to report on a scale of 0-9 how certain they were that they selected the first-stage state that would lead to the best outcome and to estimate the number of outcome units they thought they would receive at the second-stage on a scale of 0-9. Certainty and outcome estimates were not elicited on the same trial.

Anxiety. As preregistered, anxiety was measured using the State-Trait Anxiety Inventory (STAI) State Anxiety subscale.[23] STAI is a 20-item measure on a 4-point scale ranging from "almost never" to "almost always". Trait anxiety is defined as general subjective feelings of apprehension, tension, nervousness, worry, and activation/arousal of the autonomic nervous system. Internal consistency for anxiety in this sample was Cronbach's α=.94. Individual differences were analyzed with respect to difference in model-based control and learning rate across task variants operationalized as each model parameter for the protection variant minus the corresponding parameter for the non-protection variant within a given study. Positive values reflect increased model-based control and learning rates for the protection variant.

Preregistration. The main hypotheses and methods were preregistered on the Open Science Framework (OSF), https://osf.io/4j3qz/registrations.

Data availability. Task code and raw data are available through OSF, https://osf.io/4j3qz/.

Acknowledgements

DM and SMT are supported by the US National Institute of Mental Health grant no. 2P50MH094258 and Templeton Foundation grant TWCF0366. TW is supported by a Professor Anthony Mellows Fellowship. We thank Alexandra Hummel for her help with task development.

1. Tashjian, S. M., Zbozinek, T. D. & Mobbs, D. A Decision Architecture for Safety Computations. *Trends in Cognitive Sciences* **25**, 342–354 (2021).
2. Dayan, P. & Daw, N. D. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience* **8**, 429–453 (2008).
3. Dolan, R. J. & Dayan, P. Goals and Habits in the Brain. *Neuron* **80**, 312–325 (2013).
4. Drummond, N. & Niv, Y. Model-based decision making and model-free learning. *Current Biology* **30**, R860–R865 (2020).
5. Kim, D., Park, G. Y., O′Doherty, J. P. & Lee, S. W. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nat Commun* **10**, 5738 (2019).
6. Decker, J. H., Otto, A. R., Daw, N. D. & Hartley, C. A. From Creatures of Habit to Goal-Directed Learners. *Psychol Sci* **27**, 848–858 (2016).
7. Voon, V. *et al.* Disorders of compulsivity: a common bias towards learning habits. *Mol Psychiatry* **20**, 345–352 (2015).
8. Guitart-Masip, M., Duzel, E., Dolan, R. & Dayan, P. Action versus valence in decision making. *Trends in Cognitive Sciences* **18**, 194–202 (2014).
9. Wang, O., Lee, S. W., O'Doherty, J., Seymour, B. & Yoshida, W. Model-based and model-free pain avoidance learning. *Brain and Neuroscience Advances* **2**, 239821281877296 (2018).
10. Wise, T. & Dolan, R. J. Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications* **11**, 4179 (2020).
11. Elliot, A. J. The Hierarchical Model of Approach-Avoidance Motivation. *Motiv Emot* **30**, 111–116 (2006).

12. Guitart-Masip, M. *et al.* Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage* **62**, 154–166 (2012).
13. Worbe, Y. *et al.* Valence-dependent influence of serotonin depletion on model-based choice strategy. *Mol Psychiatry* **21**, 624–629 (2016).
14. Alves, H., Koch, A. & Unkelbach, C. Why Good Is More Alike Than Bad: Processing Implications. *Trends in Cognitive Sciences* **21**, 69–79 (2017).
15. Dayan, P., Niv, Y., Seymour, B. & D. Daw, N. The misbehavior of value and the discipline of the will. *Neural Networks* **19**, 1153–1160 (2006).
16. Mobbs, D., Headley, D. B., Ding, W. & Dayan, P. Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences* **24**, 228–241 (2020).
17. Tully, S., Wells, A. & Morrison, A. P. An exploration of the relationship between use of safety-seeking behaviours and psychosis: A systematic review and meta-analysis. *Clinical Psychology & Psychotherapy* **24**, 1384–1405 (2017).
18. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychol Sci* **28**, 1321–1333 (2017).
19. Sebold, M. *et al.* Reward and avoidance learning in the context of aversive environments and possible implications for depressive symptoms. *Psychopharmacology* **236**, 2437–2449 (2019).
20. Patzelt, E. H., Kool, W., Millner, A. J. & Gershman, S. J. Incentives Boost Model-Based Control Across a Range of Severity on Several Psychiatric Constructs. *Biological Psychiatry* **85**, 425–433 (2019).
21. Kool, W., Cushman, F. A. & Gershman, S. J. When Does Model-Based Control Pay Off? *PLoS Comput Biol* **12**, (2016).
22. Kurdi, B., Gershman, S. J. & Banaji, M. R. Model-free and model-based learning processes in the updating of explicit and implicit evaluations. *PNAS* **116**, 6035–6044 (2019).

23. Spielberger, C. D. *Manual for the State-Trait Inventory STAI (Form Y)*. (Mind Garden, 1983).
24. Corr, P. J. Approach and Avoidance Behaviour: Multiple Systems and their Interactions. *Emotion Review* **5**, 285–290 (2013).
25. Lebreton, M., Bacily, K., Palminteri, S. & Engelmann, J. B. Contextual influence on confidence judgments in human reinforcement learning. *PLoS Comput Biol* **15**, e1006973 (2019).
26. Ershadmanesh, S., Miandari, M., Vahabie, A. & Ahmadabadi, M. N. Higher Meta-cognitive Ability Predicts Less Reliance on Over Confident Habitual Learning System. *bioRxiv* 650556 (2019) doi:10.1101/650556.
27. Duits, P. *et al.* Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depress Anxiety* **32**, 239–253 (2015).
28. Loerinc, A. G. *et al.* Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clin Psychol Rev* **42**, 72–82 (2015).
29. Lohr, J. M., Olatunji, B. O. & Sawchuk, C. N. A functional analysis of danger and safety signals in anxiety disorders. *Clinical Psychology Review* **27**, 114–126 (2007).
30. Sangha, S., Diehl, M. M., Bergstrom, H. C. & Drew, M. R. Know safety, no fear. *Neuroscience & Biobehavioral Reviews* **108**, 218–230 (2020).
31. Helbig-Lang, S. & Petermann, F. Tolerate or Eliminate? A Systematic Review on the Effects of Safety Behavior Across Anxiety Disorders. *Clinical Psychology: Science and Practice* **17**, 218–233 (2010).
32. Odriozola, P. & Gee, D. G. Learning About Safety: Conditioned Inhibition as a Novel Approach to Fear Reduction Targeting the Developing Brain. *AJP* appi.ajp.2020.20020232 (2020) doi:10.1176/appi.ajp.2020.20020232.

33. Meyer, H. C. *et al.* Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. *PNAS* **116**, 26970–26979 (2019).
34. Thwaites, R. & Freeston, M. H. Safety-Seeking Behaviours: Fact or Function? How Can We Clinically Differentiate Between Safety Behaviours and Adaptive Coping Strategies Across Anxiety Disorders? *Behavioural and Cognitive Psychotherapy* **33**, 177–188 (2005).
35. Morriss, J., Zuj, D. V. & Mertens, G. The role of intolerance of uncertainty in classical threat conditioning: Recent developments and directions for future research. *International Journal of Psychophysiology* **166**, 116–126 (2021).
36. Morriss, J., Wake, S., Elizabeth, C. & van Reekum, C. M. I Doubt It Is Safe: A Meta-analysis of Self-reported Intolerance of Uncertainty and Threat Extinction Training. *Biological Psychiatry Global Open Science* **1**, 171–179 (2021).
37. Tanovic, E., Gee, D. G. & Joormann, J. Intolerance of uncertainty: Neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clin Psychol Rev* **60**, 87–99 (2018).
38. Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* **70**, 153–163 (2017).
39. Filipowicz, A. L. S., Levine, J., Piasini, E., Tavoni, G. & Gold, J. I. The complexity of model-free and model-based learning strategies. 30.
40. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, 1–32 (2017).
41. Watanabe, S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research* **11**, 3571–3594 (2010).

42. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
43. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface* **14**, 20170213 (2017).

Figures.

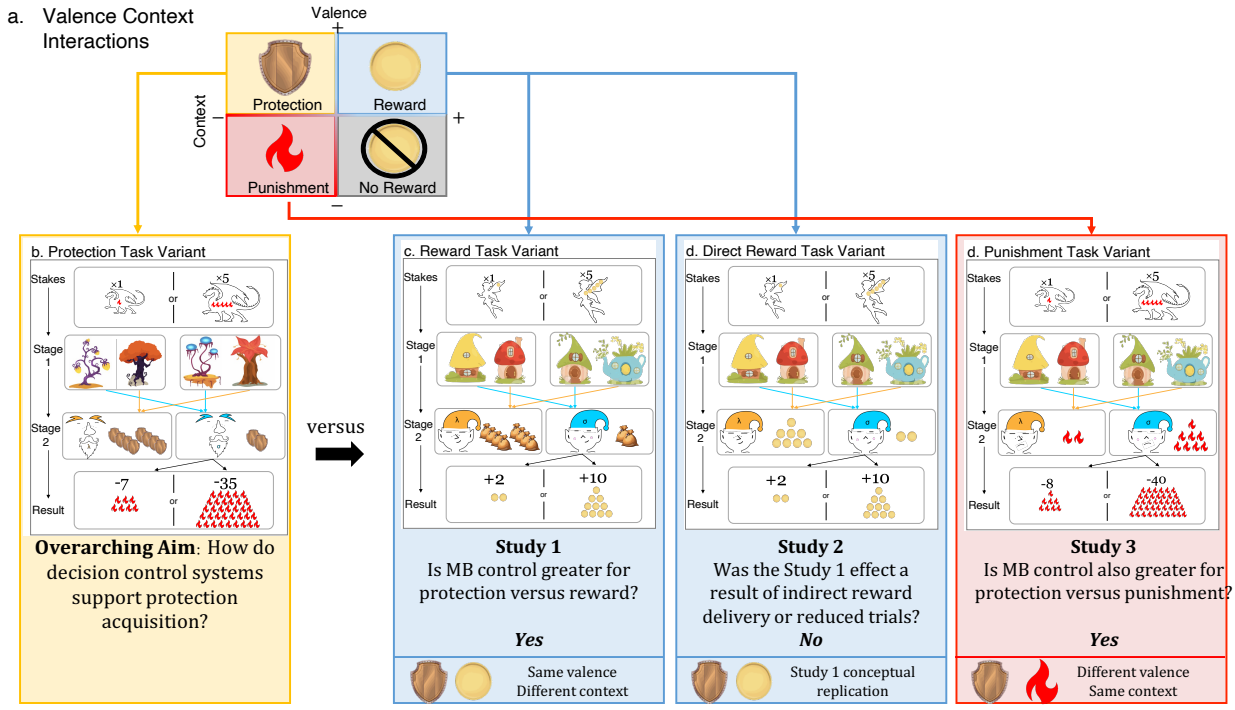
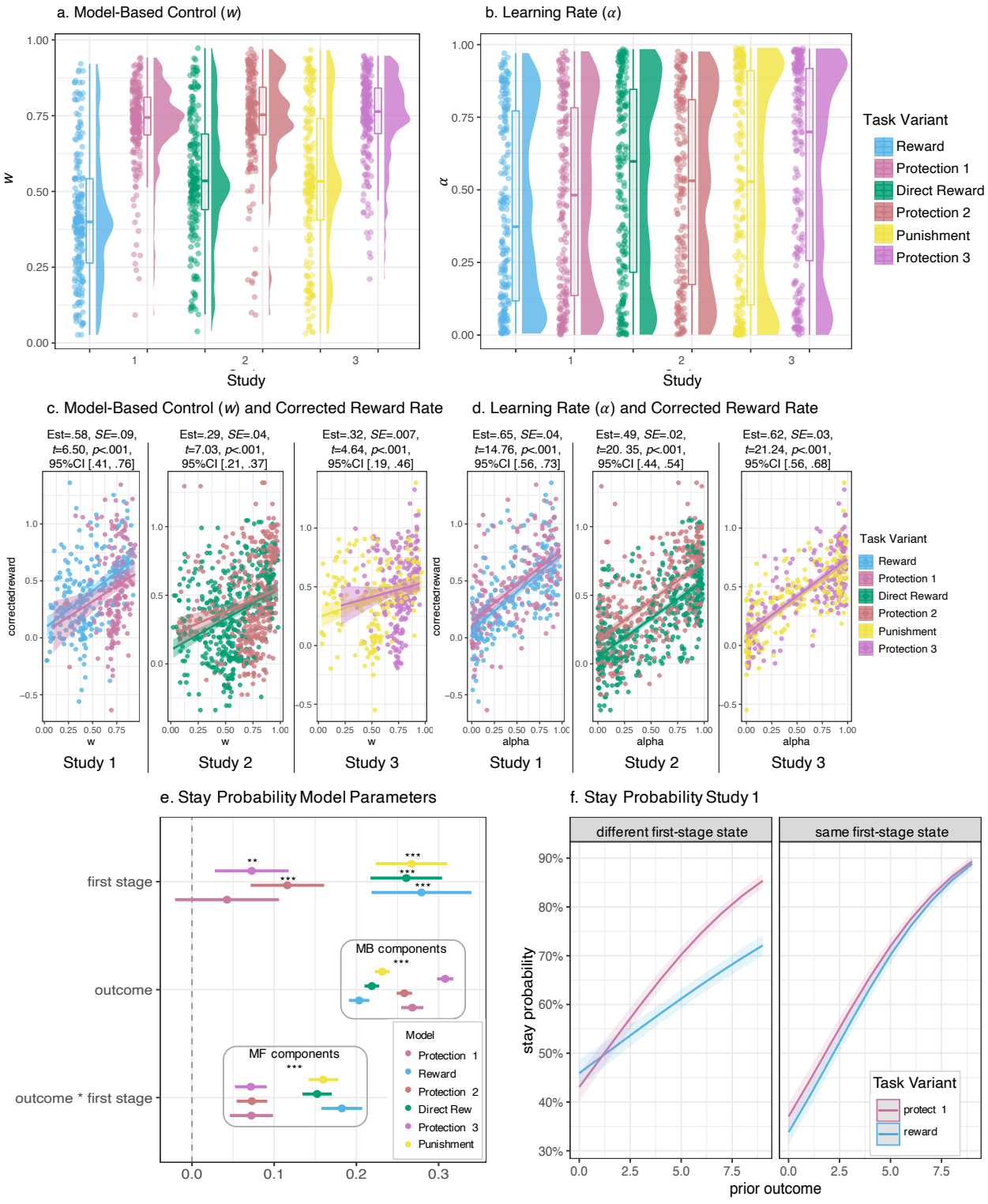


Fig. 1. Study structure. **(a)** Protection acquisition shares positive valence features with appetitive reward and negative context features with aversive punishment. The context-valence asymmetry of protection acquisition was hypothesized to be reflected in distinct engagement of decision control systems compared with stimuli in consistently appetitive or aversive domains. All studies included the **(b)** protection acquisition task variant and a comparison task variant: **(c)** reward acquisition in Study 1, **(d)** direct reward acquisition in Study 2, and **(e)** punishment avoidance in Study 3. Study 1 compared protection and reward (b versus c) using abbreviated task versions comprised of 100 non-practice trials. Study 2 compared protection and direct reward (b versus d) using longer task versions comprised of 200 non-practice trials. Study 3 compared protection and punishment (b versus e) using the longer task versions comprised of 200 non-practice trials. Deterministic transition structures are depicted with blue and orange arrows and indicate that the same first-stage state always leads to the same second-stage state. At the start of each trial, subjects saw the stakes amplifier, which showed “x1” for low-stake trials or “x5” for high-stake trials. Low-stakes results ranged from 0-9 units whereas high-stakes results ranged from 0-45 units. The stakes amplifier was applied to the punishment/reward available on that trial as well as the outcomes received. Next, subjects saw one of two pairs of first-stage dwellings (e.g., trees or houses). After subjects chose between the left and right dwelling depicted, they transitioned to the second-stage creature (e.g., gnomes or elves). Second-stage creatures delivered outcomes in the form of shields (protection), sacks (reward), coins (direct reward), or flames (punishment). At the second-stage, subjects received outcomes ranging between 0-9 according to a drifting outcome rate. Outcomes changed slowly over the course of the task according to independent Gaussian random walks (σ=2) with reflecting bounds at 0 and 9 to encourage learning throughout. Outcomes were multiplied by stakes and presented as final results applied to the maximum reward/penalty available on each trial. For example, in panel (a), subjects visited the low-payoff second-stage gnome. This gnome delivered two shields. When two shields were delivered on a low-stakes trial, which had the threat of 9 dragon flames, the end result was 2 flames (9 minus 7). When two shields were

delivered on a high-stakes trial, which had a threat of 45 dragon flames (9 flames multiplied by the stakes amplifier of 5), the end result was 35 flames (45 minus 10, 10 is calculated from 2 shields multiplied by the stakes amplifier of 5).


a. Model-Based Control (w)
b. Learning Rate (α)
Task Variant
Reward
Protection 1
Direct Reward
Protection 2
Punishment
Protection 3
Study
c. Model-Based Control (w) and Corrected Reward Rate
d. Learning Rate (α) and Corrected Reward Rate
Est=.58, SE=.09, t=6.50, p<.001, 95%CI [.41, .76]
Est=.29, SE=.04, t=7.03, p<.001, 95%CI [.21, .37]
Est=.32, SE=.007, t=4.64, p<.001, 95%CI [.19, .46]
Est=.65, SE=.04, t=14.76, p<.001, 95%CI [.56, .73]
Est=.49, SE=.02, t=20. 35, p<.001, 95%CI [.44, .54]
Est=.62, SE=.03, t=21.24, p<.001, 95%CI [.56, .68]
correctedreward
w
alpha
Study 1
Study 2
Study 3
e. Stay Probability Model Parameters
first stage
outcome
outcome * first stage
MB components
MF components
Model
Protection 1
Reward
Protection 2
Direct Rew
Protection 3
Punishment
f. Stay Probability Study 1
different first-stage state
same first-stage state
stay probability
prior outcome
protect 1
reward

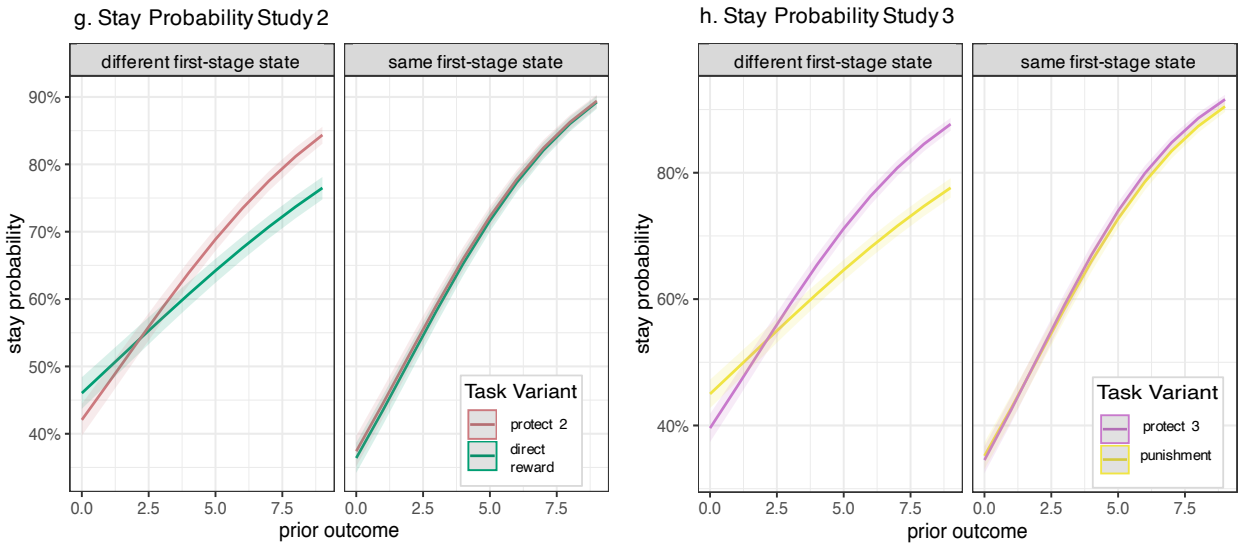


Fig. 2. Model-Based Control and Learning Rate Results. **a-b.** Raincloud plots depicting model-based control weighting (ω) and learning rate (α) by study and task variant. ω was significantly higher for the protection variants compared to all other task variants. α did not significantly differ across task variants. Far right legend indicates task variants across all studies: Study 1 = Reward and Protection 1, Study 2 = Direct Reward and Protection 2, Study 3 = Punishment and Protection 3. **c-d.** Scatterplots and linear regression lines depicting positive associations between both ω and α with corrected reward rate by study and task variant. Higher ω and α were significantly associated with higher corrected reward rate for all task variants. **e.** Mixed-effects model parameters testing contributions of the first-stage state, prior trial outcome, and interaction between first-stage state and prior trial outcome on stay probabilities. Effects from both model-free and model-based contributions were observed. MF=model-free control; MB=model-based control. **f-h.** Mixed-effects models testing model-based (different first-stage state) and model-free (same first-stage state) contributions to stay probabilities (likelihood of repeating the same second-stage state). Increased model-based contributions were revealed on the protection task variants compared with all other task variants.

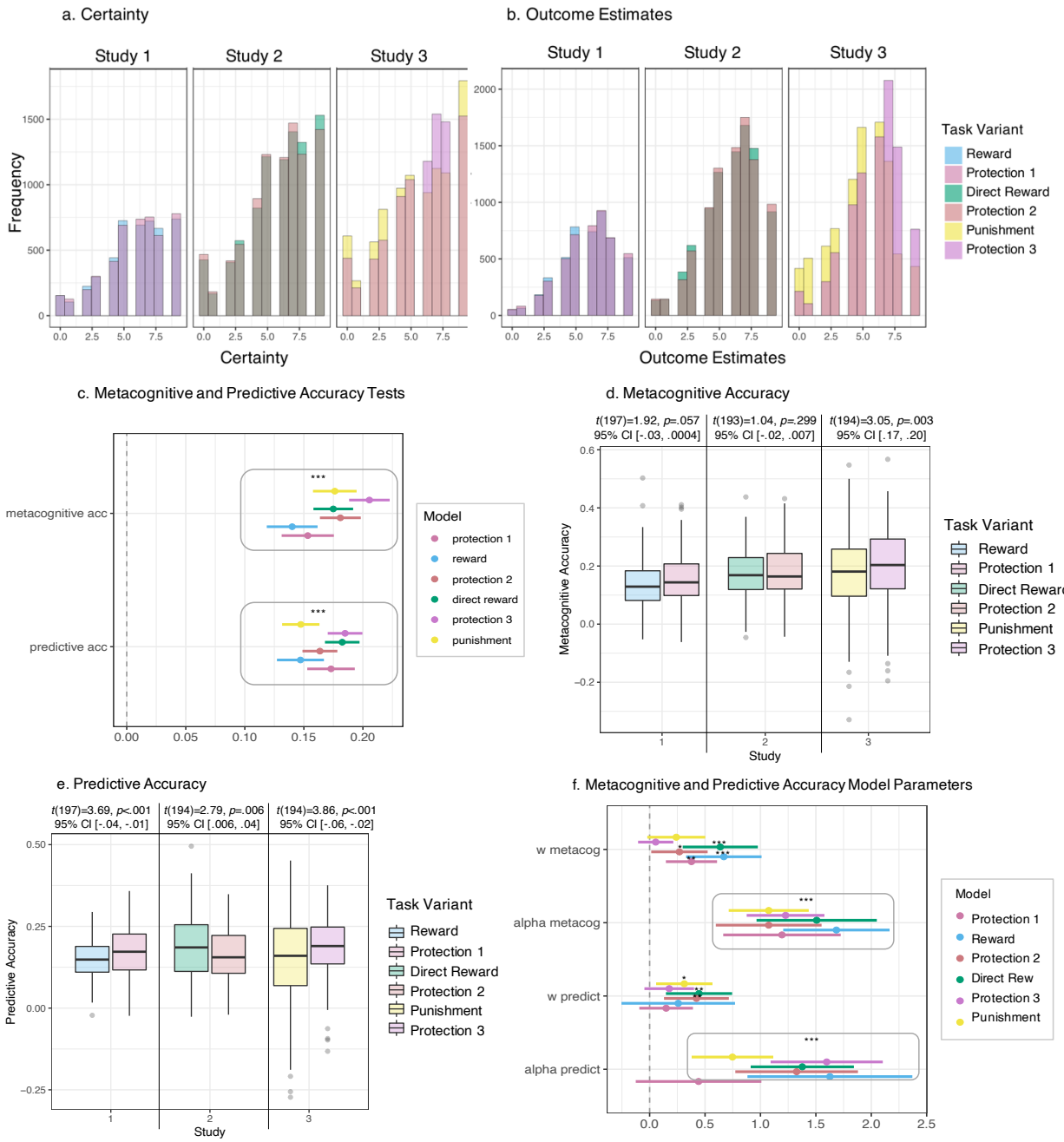


Fig. 3. Metacognitive and Predictive Accuracy Results. **a.** Histograms depicting Certainty ratings by study and task variant. Certainty was rated with respect to how sure subjects felt they were that they selected the first-stage state that would lead to the most optimal outcomes. Certainty ratings were made on a scale of 0-9 from not at all certain to very certain. **b.** Histograms depicting Outcome Estimates by study and task variant. Outcome Estimates were provided with respect to how many outcome units subjects thought they would receive at the second-stage. Outcome Estimates were made on a scale of 0-9 outcome units (i.e., subjects who rated a 2 thought they would receive 2 shields/sacks/coins/flames, respectively). **c.** Mixed-effects model parameters testing metacognitive and predictive accuracy by modeling actual outcome received as a function of Certainty and Outcome Estimates, respectively. **d.** Metacognitive accuracy boxplots by study and task variant. Metacognitive accuracy was calculated by extracting random slope coefficients

from the model of outcome predicting Certainty. Significant differences were only identified in Study 3 with higher accuracy for the protection acquisition variant compared to the punishment avoidance variant. **e.** Predictive accuracy boxplots by study and task variant. Predictive accuracy was calculated by extracting random slope coefficients from the model of outcome predicting Outcome Estimates. Significantly higher accuracy was revealed for the protection acquisition variants compared to the reward acquisition and punishment avoidance variants, but not compared to the direct reward variant. **f.** Model parameters for metacognitive and predictive accuracy coefficients regressed against model-based control weighting (ω) and learning rate (α) parameters for each task variant. Far right legends indicate task variants across all studies: Study 1 = Reward and Protection 1, Study 2 = Direct Reward and Protection 2, Study 3 = Punishment and Protection 3.

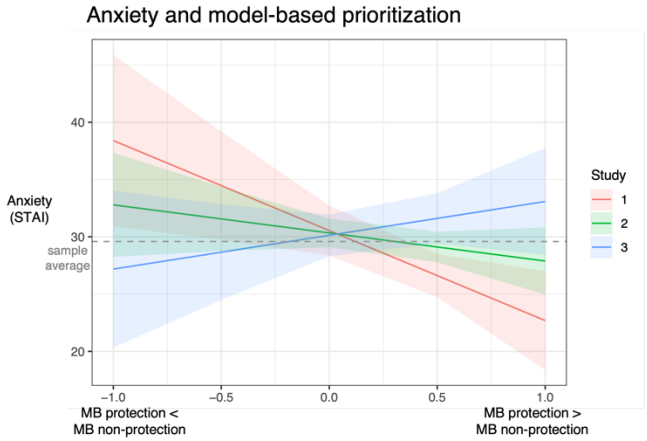


Fig. 4. Anxiety and model-based weighting (ω) by Study. Model-based prioritization was observed for protection compared with punishment avoidance (Study 3) for individuals with higher anxiety scores and for protection compared with reward acquisition for individuals with lower anxiety scores (Study 1 and 2). Dashed grey line represents the sample average scores on the State-Trait Anxiety Inventory (STAI) Trait Anxiety subscale.

SUPPLEMENTAL MATERIALS

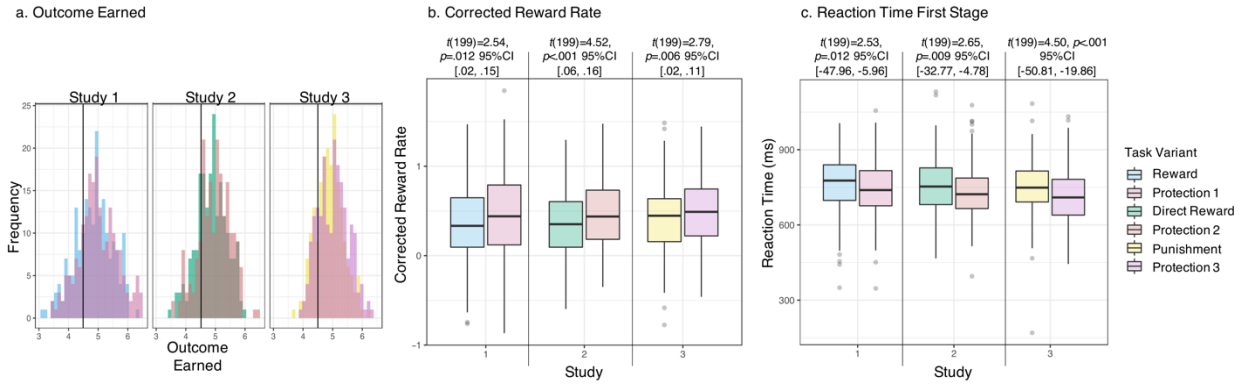


Fig. S1. Task Performance. **a.** Histograms depicting the number of outcome units earned by study and task variant. Black lines indicate median available outcome for each study. **b.** Corrected reward rate boxplots by study and task variant. Corrected reward rate was significantly higher for the protection task variants compared to all other task variants. Corrected reward rate was calculated as the average outcome earned divided by average outcome available, which was determined by the randomly drifting outcome distributions generated for each subject. **c.** Reaction time (milliseconds, ms) boxplots by study and task variant. Subjects made first-stage decisions quicker for the protection task variants compared to all other task variants. Far right legend indicates task variants across all studies: Study 1 = Reward and Protection 1, Study 2 = Direct Reward and Protection 2, Study 3 = Punishment and Protection 3.

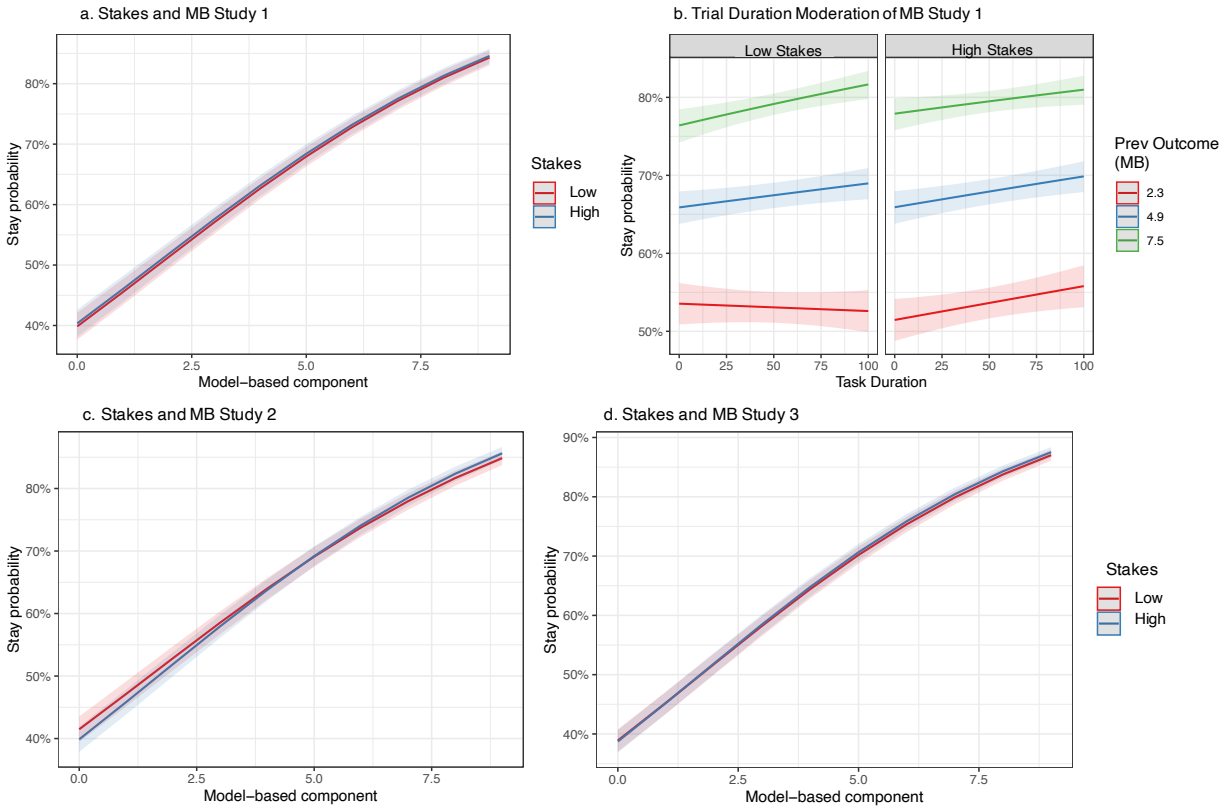


Fig. S2. We assessed whether use of model-based control was affected by stakes by testing whether stakes moderated stay behavior in mixed model analyses. **(a)** Stakes did not significantly interact with either the model-based or model-free component across tasks in Study 1: MB Estimate=.0001, SE=.009, z=.02, p=.988, 95% CI [-.02, .02], τ00=.48, R2=.15; MF Estimate=-.02, SE=.02, z=.89, p=.375, 95% CI [-.02, .05], τ00=.49, R2=.17. **(b)** Task duration interacted with stakes and previous outcome, such that there was no effect of stakes at the start of the task but high-stakes trials had an increase in likelihood of stay behavior at the end of the task: Estimate=-.001, *SE*=.0003, *z*=-2.10, *p*=.036, 95% CI [-.001, -.00004], τ_{00}=.48, R^2=.16. **(c)** Study 2, which increased trials to 200 non-practice (compared with 100 non-practice in Study 1), which revealed a stakes effect that interacted with the model-based component: Estimate=.01, *SE*=.006, *z*=2.20, *p*=.028, 95% CI [.002, .03], τ_{00}=.51, R^2=.16. This effect was driven by the direct reward variant: direct reward Estimate=.02, *SE*=.009, *z*=2.68, *p*=.007, 95% CI [.006, .04], τ_{00}=.59, R^2=.17; protection Estimate=.004, *SE*=.009, *z*=.37, *p*=.711, 95% CI [-.02, .02], τ_{00}=.66, R^2=.21. No significant interaction was present for the model-free component: Estimate=-.007, *SE*=.01, *z*=-.53, *p*=.595, 95% CI [-.03, .02], τ_{00}=.51, R^2=.17. **(d)** In Study 3, the stakes effect was not significant with respect to either model-based or model-free component: MB Estimate=.006, *SE*=.007, *z*=.88, *p*=.381, 95% CI [-.01, .02], τ_{00}=.46, R^2=.17; MF Estimate=.02, *SE*=.01, *z*=1.87, *p*=.062, 95% CI [-.001, .05], τ_{00}=.46, R^2=.18.

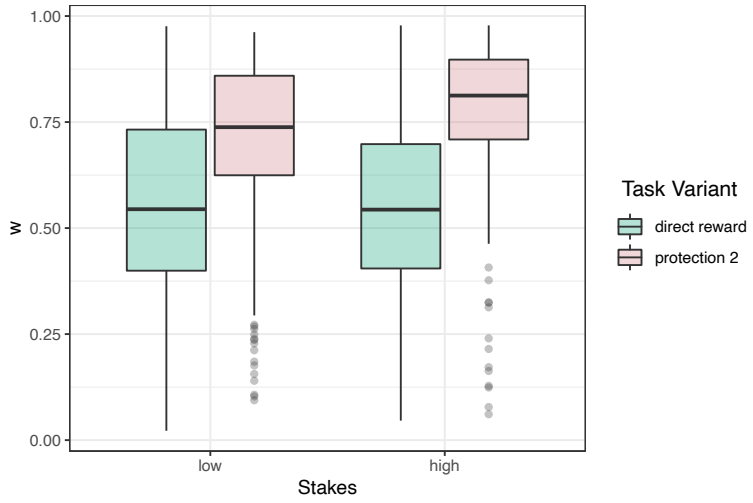


Fig. S3. Study 2 revealed a stakes effect such that model-based weighting differed between tasks for both high and low stakes, with the protection variant demonstrating more model-based control for both stakes: high stakes $w_{protection}$=.70(.21), $w_{direct.reward}$=.55(.23), *t*(199)=7.40, *p*<.001, 95% CI [.11, .19], low stakes $w_{protection}$=.77(.19), $w_{direct.reward}$=.55(.23), *t*(199)=11.82, *p*<.001, 95% CI [.18, .26].

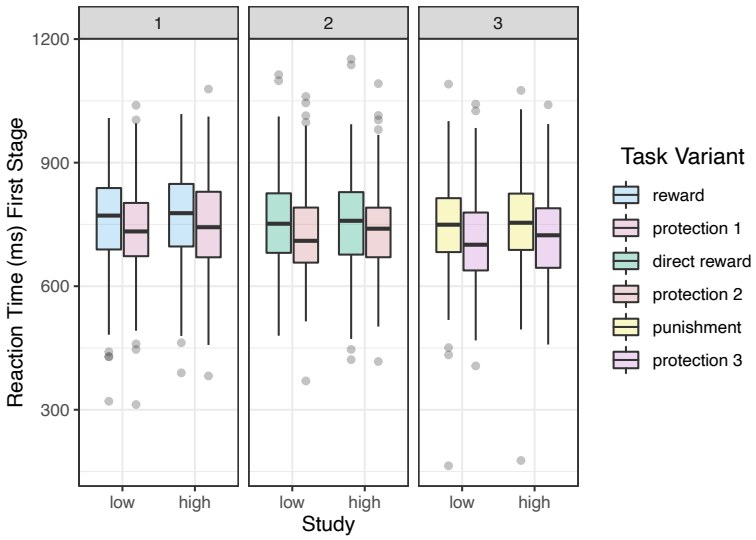


Fig. S4. RT for first-stage decisions only differed by stakes for Study 3, such that RTs were slower for high stakes: Estimate=9.84, *SE*=4.75, *t*=2.07, *p*=.039, 95% CI [.52, 19.16], σ^2=67.23, τ_{00}=85.23, R^2=.63.

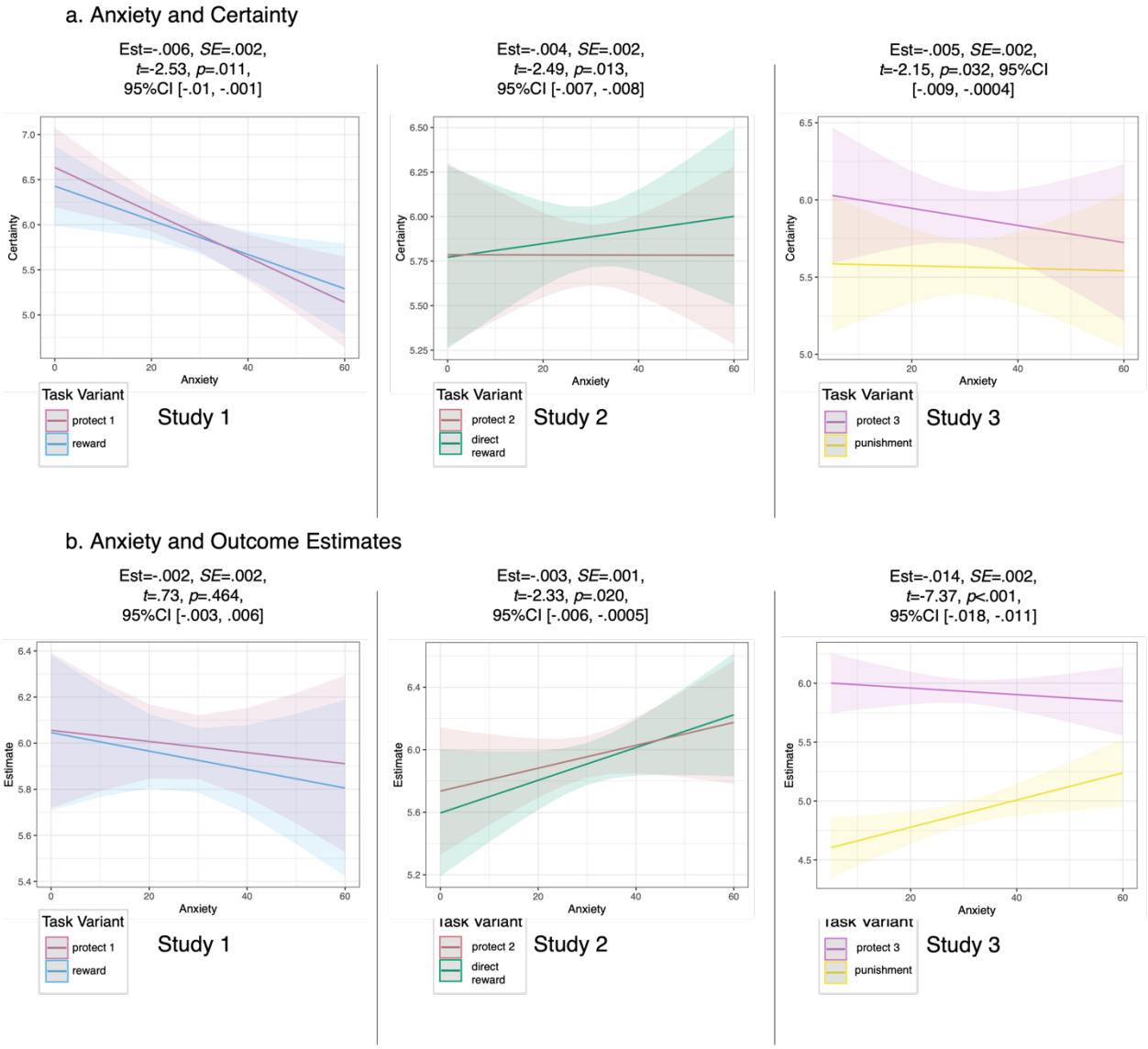


Fig S5. (a) Effects on Certainty of the interaction between anxiety (STAI) and task-variant by Study. **(b)** Effects on Outcome Estimates of the interaction between anxiety (STAI) and task-variant by Study.

Table S1. Eligibility trace (λ), stickiness (π), and inverse-temperature (β) parameters by study for the best fitting model in each study.

Study	λ	π	β
1	.80(.06)	.53(.21)	.32(.19)
2	.79(.12)	.51(.29)	.51(.23)
3	.86(.10)	.51(.24)	.50(.24)

Note. Values are average across all subjects with standard deviation in parenthesis.