Human hacks and bugs in the recruitment of reward systems for goal achievement

Gaia Molinaro (gaiamolinaro@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA, 94704 United States

Anne G. E. Collins (annecollins@berkeley.edu)
Department of Psychology & Helen Wills Neuroscience Institute, University of California, Berkeley
Berkeley, CA, 94704 United States

Abstract

Human learning is often motivated by self-imposed challenges, which guide behavior even in the absence of external rewards. Previous studies have shown that humans can use personal goals to "hack" the definition of reward, warranting an extension of the classic reinforcement learning framework to account for the flexible attribution of value to outcomes according to current goals. However, learning through goal-derived outcomes is less efficient than learning through more established reinforcers, such as numeric points. At least three possible explanations exist for this sort of impairment, or "bug". First, occasional lapses in executive function, which is required to encode and recognize goals, may result in subsequent failure to update values accordingly. Second, the higher working memory load required to encode novel stimuli as desirable outcomes may impair people's ability to update and remember correct stimulus-reward associations. Third, a weaker commitment to arbitrary goals may result in dimmer appetitive signals. By extending existing experimental paradigms that include learning from both familiar rewards and abstract, goal-contingent outcomes and combining them with computational modeling techniques, we find evidence for each of the proposed accounts. While other factors might also play a role in this process, our results provide an initial indication of the key elements supporting (or impairing) the attribution of rewarding properties to otherwise neutral stimuli, which enable humans to better pursue arbitrarily set goals.

Keywords: human reinforcement learning; goal-directed learning; working memory; computational cognitive modeling

Introduction

The reinforcement learning (RL) framework has proven useful in explaining neural and behavioral signatures of learning in natural organisms, including humans (Niv, 2009). RL also plays a key role in many of the recent advances in artificial intelligence — sitting at the root of self-driving cars and machines that can beat human Go masters (Silver et al., 2016; Wang et al., 2018). However, even the most sophisticated RL algorithms are strikingly inflexible compared to humans. In particular, standard RL systems depend on external rewards to learn how to select the optimal action given a certain state of the environment. For software and robots, rewards are usually defined by the engineer in alignment with a specific objective the RL agent is designed to attain (Sutton & Barto, 2018). Biological rewards are shaped by evolution or learned through association with homeostatic or reproductive rewards. However, humans can also subjectively attribute rewarding properties to virtually any otherwise neutral state of the world, thereby "hacking" the innate reward system to accomplish arbitrarily set goals (abstract representations of future states one is trying to bring about; Juechems & Summerfield, 2019; O'Reilly, 2020). Testifying to the powerful role of personal goals in the definition of reward, people frequently incur physical or cognitive costs for the pursuit of feats that are more or less far removed from evolutionary advantages (e.g., running a marathon or completing crosswords; Blain & Sharot, 2021). The ability to attribute value to intermediary goals is also at the core of "divide and conquer" strategies, which are required to accomplish complex behavior (Diuk et al., 2013; Newell & Simon, 1972).

Upon investigating this phenomenon more closely in a probabilistic learning task, McDougle et al. (2022) observed similar RL-related signals in dopaminergic structures upon receipt of both standard reinforcers (numeric points) and goal outcomes (fractal images associated with goal-congruent). Nonetheless, learning was less effective when feedback was provided as novel goal-contingent rewards than as standard reinforcers, damaging performance. Here, we replicate evidence for the human ability to "hack" the reward system to achieve arbitrarily set goals while attempting to explain the sources of imperfection ("bugs") in this process.

McDougle et al. (2022) observed that stronger prefrontal engagement at goal encoding leads to more reliable reward signals when receiving goal-contingent feedback. Therefore, one possible account for worse performance in the instructed rewards condition is that occasional lapses in goal maintenance could subsequently prevent the recognition of goal attainment (or failure to do so). Here, we test this hypothesis by adapting previous experiments and directly measuring people's encoding of abstract feedback as either rewarding or non-rewarding (Experiment 1).

Needing to imbue different abstract goals with value on every trial, as was the case in previous experiments, could also decrease participants' ability to recruit resource-limited working memory to guide choice (Collins & Frank, 2012). Here, we eliminate this confound by implementing a control task in which goal-contingent outcomes remained stable throughout the experiment (Experiment 2).

A third, non-exclusive explanation for lower performance when learning from arbitrary goals than numeric points relates to differences in reward-related signals between the two. While numeric points were likely associated with positive outcomes on multiple occasions prior to participants' engagement with our task, goal-contingent outcomes were com-

pletely abstract and instructed by the experimenter. Moreover, greater familiarity and repeated associations with reward might have made points more immediately rewarding than goal images. Thus, it is possible that commitment to accrue points was stronger than the willingness to obtain goal-congruent images, which may have resulted in weaker appetitive responses to the latter, compared to the former. This hypothesis is further justified by the fact that, in the McDougle et al. (2022) study, reward prediction error signals following goal-contingent outcomes failed to reach statistical significance in the dorsomedial striatum, despite there not being differences between conditions in whole-brain contrasts. Here, we employ behavioral and computational analyses across experiments to test this hypothesis.

Experimental Design

The experimental design was adapted from Collins and Frank (2012) and McDougle et al. (2022) with the aim of comparing learning performance in tasks where feedback was provided in the form of either points (standard reinforcers) or abstract, novel stimuli associated with instructed goals (one-shot encoded, goal-dependent rewards). Two versions of the experiment ("Experiment 1" and "Experiment 2" below) were run.

In both experiments, participants were presented with thorough instructions and a round of practice trials, and then tasked with learning the correct mapping between each of six images and one of three actions (Figure 1). The images belonged to the same category (e.g., vegetables, cartoon characters, famous monuments) and were unique to each of six blocks. Participants were told that learning the correct response for one image was not informative with respect to the correct action for another image. Within each block, individual images were presented 13-14 times with a uniform delay between same-stimulus presentations. Three "Points" and three "Goals" blocks were pseudo-randomly interleaved.

Within Points blocks, participants first saw a "+1" and a "+0" message at the top and bottom of the screen, with the labels "Win" and "Lose" at the top of each respective message, identifying the desirable outcome. Next, participants viewed a stimulus image and had to press one of the J, K, or L keys on their keyboard. Depending on whether their response was correct, they received feedback in the form of the +1 message or +0 message in a deterministic fashion and without the associated label. The structure in Goals blocks was similar, but +1 and +0 messages were replaced by fractal images, accompanied by the labels "Goal" at the top, and "Nongoal" at the bottom of the screen. Feedback in the Goals blocks was presented in the form of the goal or nongoal image (according to whether the participant selected the correct key) without the associated label. To track any occurring lapses in the recognition of desirable outcomes, participants were instructed to "collect" +1 messages and goal images by pressing a separate key (D) upon obtaining them, while avoiding the collection of +0 messages and nongoal images. If participants collected the outcome within the allotted time, the black square that originally surrounded the outcome would turn blue, signaling that their "collection" was recorded (but not whether it was correct).

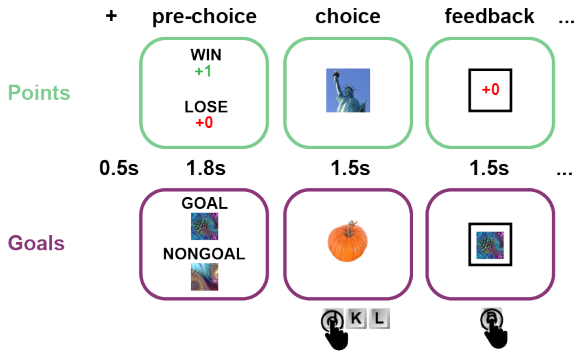

Figure 1: Task design. Points and Goals trials were presented in separate blocks (three each) in a pseudo-randomized order. In Experiment 1, "Goal" and "Nongoal" images changed on each trial. In Experiment 2, the same two fractals were used as goal and nongoal images.

In Experiment 1, goal and nongoal images changed on every trial. In Experiment 2, the same two fractals were used as goal and nongoal images (randomized across participants while maximizing visual distinctiveness) throughout the experiment. Participants in Experiment 2 were shown the goal and nongoal images during the instructions. In Experiment 2, participants also rated their liking of the goal and nongoal images on a scale from 1 ("not at all") to 5 ("very much") at three different time points: 1) before the instructions began, 2) after being told which images would be used as the goal and nongoal image throughout the experiment, 3) at the end of the experiment. Experiment 1 and Experiment 2 were otherwise identical.

After completing the main task, participants performed six blocks of an N-back task (Kirchner, 1958), which was intended to measure working memory independently of the learning task (Haatveit et al., 2010). Here, participants viewed a sequence of letters and were tasked with pressing the left arrow key on their keyboard when the same letter reappeared following the presentation of N stimuli. The relevant N (1–3) was specified at the beginning of the block.

Participants

All subjects were recruited from the university's pool of participants and completed the experiment online on their own devices. Participants received partial course credit as compensation for their time.

One hundred and twenty-five individuals completed Experiment 1. Thirty-five participants were excluded because they met at least one of the following criteria: 1) missing more than 25 trials during the course of the experiment, 2) missing more than 10 trials consecutively, 3) selecting the same response more than 15 times in a row, 4) having a reward

collection *d'* in the Points condition below or equal to 1, 5) making more than 20 reward collection errors in the Points condition, 6) having a reward collection *d'* in the Goals condition below or equal to 0, 7) making more than 50 reward collection errors in the Goals condition. These quality checks were intended to control for task engagement, and thresholds were based on elbow points in ordered group distributions (Xia et al., 2021). Therefore, data from 90 participants (70% female, ages 18-30, M = 20.33 $\pm$ 0.21) were analyzed for Experiment 1.

One hundred and twenty-one individuals completed Experiment 2. Twenty-eight participants were excluded based on the same criteria as in Experiment 1. Therefore, data from 93 participants were analyzed for Experiment 2 (71% female, ages 18-27, M = 20.09 $\pm$ 0.19).

The University of California, Berkeley Institutional Review Board approved the experimental procedure.

Computational Modeling

We fit a set of candidate models that shared a basic RL architecture (Sutton & Barto, 2018). The value of each stimulus-action pair (Q) was initialized at 0 at the beginning of each block. Q-value updates followed the delta rule (Rescorla & Wagner, 1972), according to which the expected value of the chosen response (c) on a given trial (t) is updated as:

$$Q_{t+1}(c) = Q_t(c) + \alpha \cdot \delta_{c,t} \quad (1)$$

where α is the learning rate and $\delta_{c,t}$ is the trial's reward prediction error calculated upon receipt of feedback (r):

$$\delta_{c,t} = r_{c,t} - Q_t(c) \quad (2)$$

Models also included a forgetting parameter (ϕ) which partially countered feedback-guided updates by returning Q-values towards the initial value on each trial (see McDougle et al., 2022). All models selected actions via a softmax policy, with the inverse temperature parameter (β) controlling the amount of exploration of suboptimal responses. Other factors that varied across models included the presence of separate βs for Points and Goals blocks, separate αs for Points (α_P) compared to Goals (α_G) blocks, and an additional parameter, r, which rescaled the value of goal-conditioned outcomes as a fraction of 1 while leaving numeric rewards intact. Instead of objective outcomes (i.e., +1 for obtaining a point or a goal image, and +0 otherwise), a subset of the models used subjective rewards (+1 if participants pressed the "reward collection" button, +0 otherwise) to update Q-values.

All models were fit through hierarchical Bayesian modeling, using PyStan to interface the programming language Stan via Python. Weakly informative priors were chosen in accordance with Baribault and Collins (2021). Four parallel chains were run for 1000 iterations each. All models converged, as evidenced by all $\hat{R}$ scores (the Gelman-Rubin convergence diagnostic) being $\leq$ 1.01 and effective sample sizes being $\geq$ 400 (Vehtari et al., 2021). The widely applicable information criterion (wAIC; Watanabe, 2013) was used to compare the predictive value of candidate models while accounting for complexity. wAIC scores were calculated following Fontanesi et al. (2019). Model validation was performed by simulating data based on five samples from each chain's last 125 samples for each participant.

Results

Humans can attribute reinforcing properties to abstract, goal-congruent stimuli

Despite using a different experimental design, we replicated McDougle et al.'s (2022) finding that learning can be guided by a goal-contingent attribution of value to otherwise neutral stimuli, similar to learning in the presence of numeric points (which constitute more typical reinforcers). In both experiments, participants exhibited successful learning of each stimulus-action mapping, with high average accuracy in both Points (Experiment 1: M = 0.79 $\pm$ 0.01; Experiment 2: M = 0.79 $\pm$ 0.01) and Goals blocks (Experiment 1: M = 0.65 $\pm$ 0.02; Experiment 2: M = 0.76 $\pm$ 0.01; Figure 2). While not novel, this finding is remarkable, as it calls for an extension of the traditional RL framework in psychology, wherein reinforcers are typically understood as deriving from primary sources of reward or stimuli associated with them. By contrast, participants in our study were able to treat abstract stimuli as signals for learning based on a single instruction.

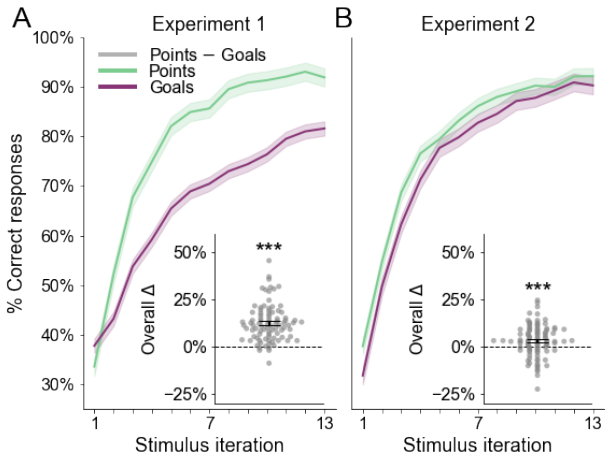

Figure 2: Learning performance in Points and Goals blocks across stimulus iterations for Experiments 1 (A) and 2 (B). Insets illustrate differences in overall performance. Shading and error bars show the standard error of the mean. Dots represent individual participants.

Learning via goal-contingent signals is less efficient than learning through standard rewards

Consistent with previous results (McDougle et al., 2022), learning was more efficient in Points blocks compared to Goals blocks (Experiment 1: $t(81) = 11.54$ $p < 0.001$; Experiment 2: $t(92) = 3.78$ $p < 0.001$). Therefore, the attainment of arbitrary goal outcomes cannot be considered fully equivalent

to receiving more familiar and easily interpretable reinforcers such as numeric outcomes. In the remainder of this paper, we test three hypotheses that might account for such differences.

Lapses in goal recognition play a marginal role in accounting for goal-dependent learning

Throughout the experiment, we recorded participants' subjective accounts of reward receipt by asking them to press a key upon obtaining points or goal images (but not other outcomes). This manipulation enabled us to directly test the hypothesis that occasional lapses in goal outcome encoding underlie differences in learning across task conditions.

Overall, reward collection errors were extremely rare. Accuracy was very high across experiments, as evidenced by high *d'* scores (a measure of sensitivity that is unaffected by response biases; Experiment 1: M = 3.61 $\pm$ 0.06; Experiment 2: M = 3.97 $\pm$ 0.05). Nonetheless, average accuracy was greater in Points than in Goals trials in Experiment 1 ($t(81) = 6.36$, $p < 0.001$). This finding is consistent with the hypothesis that setting a reward value for a new goal is more demanding for executive functions than experiencing a known reward. It also supports the view that differences in learning performance between the two conditions observed in Experiment 1 can be explained by occasional lapses in goal maintenance and subsequent recognition. This theory makes two further predictions: 1) we should observe a difference in lapses in reward collection in Experiment 2 to account for significant learning differences, and 2) differences in reward collection should predict differences in performance.

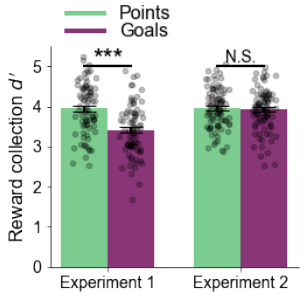

Figure 3: Overall reward collection performance, as measured by *d'*, in Points and Goals blocks for Experiments 1 and 2. *** $p < 0.001$

These predictions were not supported by the findings. Specifically, there was no significant difference in reward collection *d'* scores for Points compared to Goals trials ($t(92) = 0.43$, $p = 0.667$; Figure 3) in Experiment 2. Moreover, the difference in learning performance between the conditions was not significantly correlated with the difference in reward collection accuracy between the conditions in either experiment (Experiment 1: Spearman's $\rho = 0.03$, $p = 0.762$; Experiment 2: $\rho = 0.17$, $p = 0.133$). In addition, computationally accounting for lapses could not fully recover the difference between Points and Goals performance. This was evident in the fact that substituting objective reward contingencies with subjective outcomes (based on whether participants collected points or images) did not improve model fit (Experiment 1: ΔwAIC relative to the best model = 235.57; Experiment 2: ΔwAIC = 556.76).

Together, these results suggest that occasional lapses in goal outcome recognition may play a marginal role in the imperfect recruitment of reward structures for the achievement of arbitrary goals, but are likely not the sole cause of impairments in goal-dependent learning.

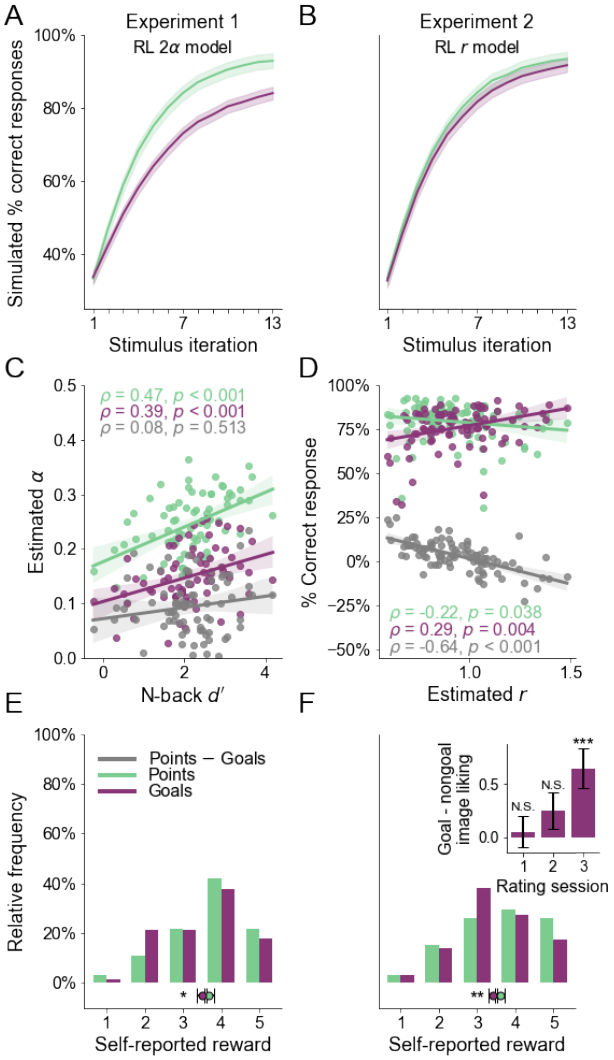

Figure 4: Computational models successfully replicate behavioral patterns in Experiments 1 (A) and 2 (B). (C) Working memory cannot fully account for differences in learning rates between conditions in Experiment 1. (D) Differences in performance between conditions correlate with individual differences in value attribution in Experiment 2. Self-reported feelings of reward were stronger for the receipt of points than "goal" images in Experiments 1 (E) and 2 (F). The inset (F) shows progressive preferential liking for "goal", compared to "nongoal" images in Experiment 2.

Working memory load impinges on goal-dependent learning

In Experiment 1, the goal image varied on each trial. This choice in experimental design pushes the concept of value attribution onto abstract novel stimuli to a logical extreme, but poses an additional load on working memory when learning from goals, compared to the Points condition (in which desirable outcomes stay identical throughout the experiment). To account for the difference in working memory load, goal images remained the same in Experiment 2 for the entire duration of the study.

Providing evidence that working memory load was an important factor in the impairment of learning with goal-dependent rewards, maintaining a fixed goal/nongoal image pair in Experiment 2 significantly reduced the difference between Points and Goals performance compared to Experiment 1 ($t(173) = 7.09$, $p < 0.001$; Figure 2). However, we note that working memory load was not the only factor impacted by changes in the task design applied in Experiment 2, and repeated exposure and value attribution to the same goal images created a possibility to improve the goal signal associated with them – a point to which we return below. In both experiments, individual differences in working memory involvement in learning were captured by the model's learning rate parameters. Indeed, in Experiment 2, the learning rate (α) was significantly correlated with N-back *d'* – a standard proxy of working memory that was measured independent of learning ($\rho = 0.35$, $p = 0.001$). In Experiment 1, the best model included separate learning rates for the two conditions, both of which correlated with N-back *d'* scores (α_P: $\rho = 0.47$, $p < 0.001$; α_G: $\rho = 0.39$, $p < 0.001$). Furthermore, the average learning rate for Points blocks was significantly higher than the learning rate for Goals blocks (difference M = 0.09, 95% HDI = [0.01, 0.19]; Figure 4B), supporting the hypothesis that working memory was less available to support learning in Goals blocks.

While taking these results into account, it is remarkable that the difference between Points and Goals trials remained significant even in Experiment 2, where we removed the working memory load confound. Indeed, the difference there was not well captured by a model with separate learning rates (ΔwAIC compared to the best model = 23.22). Moreover, even in Experiment 1, the difference between α_P and α_G was not correlated with N-back *d'* scores ($\rho = 0.08$, $p = 0.513$; Figure 4B), suggesting that differences in learning rates were at least in part accounting for factors beyond working memory. Therefore, novelty in outcome encoding appeared to have a considerable, but partial role in slowing down learning from goal-contingent outcomes compared to standard rewards.

Goal-contingent outcomes are likely associated with weaker appetitive signals

Neither cognitive explanation explored thus far could fully account for differences between learning guided by goal-contingent outcomes versus numeric points. An additional, non-mutually exclusive hypothesis of the phenomenon holds that weaker appetitive signals associated with attaining abstract outcomes, compared to the receipt of standard rewards, might have caused the learning differences observed between the two conditions.

Table 1: Logistic mixed-effects regression predicting correct versus incorrect reward collection from the (z-scored) condition, stimulus iteration, block number, and the interaction between condition and the latter two.

	Estimate $\pm$ SEM	**z**	**p**
Experiment 1			
Intercept	1.86 ± 0.01	21.11	< 0.001
Cond.	0.30 ± 0.07	4.49	< 0.001
Iteration	0.11 ± 0.01	16.63	< 0.001
Block	0.16 ± 0.02	10.32	< 0.001
Cond. $\times$ iteration	0.01 ± 0.01	1.86	0.063
Cond. $\times$ block	-0.02 ± 0.02	-1.00	0.320
Experiment 2			
Intercept	1.89 ± 0.08	22.92	< 0.001
Cond.	0.30 ± 0.07	4.48	< 0.001
Iteration	0.13 ± 0.01	18.34	< 0.001
Block	0.19 ± 0.02	12.63	< 0.001
Cond. $\times$ iteration	-0.00 ± 0.01	-0.42	0.677
Cond. $\times$ block	-0.06 ± 0.01	-3.56	< 0.001

In favor of this idea, subjective ratings of "how rewarding it felt" to obtain goal images, self-reported by participants at the end of the experiment (on a scale from 1 = "not at all" to 5 = "very much"; Experiment 1: $M = 3.49 \pm 0.14$; Experiment 2: $M = 3.41 \pm 0.11$), were lower than ratings for the attainment of points (Experiment 1: $M = 3.69 \pm 0.13$; Experiment 2: $M = 3.60 \pm 0.12$) both in Experiment 1 (Wilcoxon's $z = 172$, $p = 0.043$) and Experiment 2 ($z = 58.8$, $p = 0.010$; Figure 4E-F). Moreover, the best model in Experiment 2, where working memory load was comparable across conditions, incorporated a parameter, *r*, which multiplied goal-dependent goal signals while leaving rewards from numeric outcomes unchanged at 1. In this model, which successfully replicated behavioral patterns (Figure 4B), the mean *r* was lower than 1, although individual differences sometimes exceeded 1 ($M = 0.94$, 95% = [0.69, 1.21]), suggesting that, on average, reward signals were weaker for goal images than points. In further support of the idea that condition-based differences in value attribution caused learning differences between Points and Goals blocks, individual mean values of *r* in Experiment 2 correlated with learning performance differences between the two ($\rho = -0.64$, $p < 0.001$; Figure 4D). The same was true for *r* in Experiment 1 ($M = 0.71$, 95% HDI = [0.50, 0.93], $\rho = -0.69$, $p < 0.001$), where the model that comprised this parameter was a close second best-fitting (ΔwAIC = 13.32).

Repeated opportunities to attribute value to goal images (present in Experiment 2, but not Experiment 1) led condition-based differences in approach responses to dampen throughout the course of the experiment (as evidenced by a

significant interaction between block number and condition on reward collection accuracy; Table 1). Accordingly, in Experiment 2 participants' liking of the goal image, relative to the nongoal image, increased over the course of the experiment. If before being instructed to associate either image with desirable outcomes participants had no preference between goal and nongoal images (each measured on a scale from 1 to 5; difference M = 0.05 $\pm$ 0.15, t(92) = 0.37, p = 0.714 against 0), after completing the task, participants had acquired a significant preference for the goal image (M = 0.65 $\pm$ 0.18, t(92) =3.53, $p < 0.001$). Together, these results suggest that differences in the strength of the reward signal associated with goal-dependent outcomes, compared to numeric points, were a partial cause of learning differences between the two conditions.

Discussion

If secondary rewards (e.g., money or numeric points) earn their reinforcing property through experienced associations with primary rewards (e.g., food, water, or sex), goals can imbue even abstract and/or novel stimuli with value, enabling learning with a level of flexibility that has so far only been shown in humans. From struggling to solve a Sudoku puzzle to enduring the physical strain of climbing Kilimanjaro, people complete self-imposed challenges despite incurring varying amounts of costs. Perhaps even more strikingly, humans are capable of using goal-dependent signals to guide learning toward voluntarily set objectives – calling for an important reconsideration of how rewards are defined in RL. Here, we replicated the finding (McDougle et al., 2022) that people can attribute value to completely novel, abstract stimuli upon a single, instructed association with the task goal, and use such constructed reward signals to guide their own learning. Moreover, we confirmed previous results showing that learning from goal-contingent outcomes is less efficient than learning from more standard rewards. Understanding the origins of this discrepancy could help identify the key processes involved in flexible value attribution during goal-guided learning. We therefore sought to identify and test initial hypotheses for why this might be the case.

First, slower learning from goal-dependent outcomes could result from occasional lapses in goal-outcome maintenance and recognition. This hypothesis was justified by previous neuroimaging findings in a similar task, wherein stronger interactions between the prefrontal cortex and reward-sensitive regions predicted better performance in the Goals condition (McDougle et al., 2022). To address it, we asked participants to "collect" desirable outcomes (i.e., points or goal images), hence recording any occurring errors in the attribution of rewarding or punishing properties to the observed outcome. While reward collection errors were more frequent in Goals than Points blocks, lapses were extremely rare and unlikely to fully account for condition-based differences in learning performance.

Second, we controlled for possible effects of working memory load that might have detracted from learning resources in the original Goals condition – in which desirable and undesirable outcomes had to be newly encoded on each trial – and the Points condition – in which positive and negative outcomes remained stable over time. This manipulation significantly reduced learning differences between the two conditions, suggesting an important role of executive function in the attribution of rewarding properties to novel stimuli. At the same time, differences between points and goals-based learning were not entirely attributable to working memory limitations.

Third, we asked whether differences in learning from secondary, as opposed to goal-conditioned rewards, could be partially accounted for by differences in appetitive signals. This account was inspired by parallel differences in reward-related neural signals observed in a previous exploration (McDougle et al., 2022). Indeed, both self-reports of the subjective feeling of reward in response to goal images, compared to numeric points, and computational modeling results, provided evidence for a weaker reward-related signal in the former case. We speculate that this difference may be adaptive, preventing people from attributing too much value to arbitrary, instructed goals as opposed to more established outcomes that have been frequently associated with rewards in the past.

At this stage, we cannot confirm or exclude the role of other factors in the different effects of standard versus goal-contingent rewards. For instance, specific properties of numerical outcomes may ease the interpretation of outcomes in the Points condition (e.g., Shenhav et al., 2016). One way to test this hypothesis would be to substitute points with more abstract, yet traditionally positive/negative outcomes, such as green/red marks. Such a design may also address possible differences in the two conditions due to greater visual processing needed to encode fractals compared to numeric outcomes. Lastly, each of the factors identified here is likely to interact with the others – an aspect that awaits future research.

To summarize, we have replicated the finding that humans can imbue abstract, novel stimuli with reward and use them as signals for learning, and investigated whether occasional lapses in goal maintenance, working memory load, and differences in value attribution cause slower learning with such goal-defined reinforcers compared to standard, numeric points as rewards. Altogether, we find evidence for each of the three hypotheses, suggesting that multiple factors contribute jointly to the recruitment of the reward system while attempting to learn and attain arbitrarily set goals. While other components might be at play, the present experiments provide an initial indication of key elements that may impair flexible attribution of rewarding properties to otherwise neutral stimuli. Future studies may explore how each of these factors interacts with the others. A better understanding of the limitations of the ability to set and achieve arbitrary goals will ultimately lead to ways we can enhance this perhaps uniquely human capacity, empowering people to reach their own quotidian or ambitious goals.

Acknowledgments

We thank Beth Baribault for guidance on Bayesian modeling and Amy Zou for help setting up the experiment online. This work was supported by NSF Grant 2020844 awarded to AGEC.

References

Baribault, B., & Collins, A. (2021). Troubleshooting bayesian cognitive models: A tutorial with matstanlib.

Blain, B., & Sharot, T. (2021). Intrinsic reward: Potential cognitive and neural mechanisms. *Current Opinion in Behavioral Sciences*, *39*, 113–118.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? a behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, *33*(13), 5797–5805.

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, *26*(4), 1099–1121.

Haatveit, B. C., Sundet, K., Hugdahl, K., Ueland, T., Melle, I., & Andreassen, O. A. (2010). The validity of d prime as a working memory index: Results from the "bergen n-back task". *Journal of clinical and experimental neuropsychology*, *32*(8), 871–880.

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in cognitive sciences*, *23*(10), 836–850.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

McDougle, S. D., Ballard, I. C., Baribault, B., Bishop, S. J., & Collins, A. G. (2022). Executive function assigns value to novel goal-congruent outcomes. *Cerebral Cortex*, *32*(1), 231–247.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

O'Reilly, R. C. (2020). Unraveling the mysteries of motivation. *Trends in cognitive sciences*, *24*(6), 425–434.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Current research and theory*, 64–99.

Shenhav, A., Straccia, M. A., Botvinick, M. M., & Cohen, J. D. (2016). Dorsal anterior cingulate and ventromedial prefrontal cortex have inverse roles in both foraging and economic choice. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(6), 1127–1139.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis*, *16*(2), 667–718.

Wang, S., Jia, D., & Weng, X. (2018). Deep reinforcement learning for autonomous driving. *arXiv preprint arXiv:1811.11329*.

Watanabe, S. (2013). A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, *14*(27), 867–897.

Xia, L., Master, S. L., Eckstein, M. K., Baribault, B., Dahl, R. E., Wilbrecht, L., & Collins, A. G. E. (2021). Modeling changes in probabilistic reinforcement learning during adolescence. *PLoS computational biology*, *17*(7), e1008524.