Human goal-directed actions survive manipulations of the action-outcome contingency

Omar David Perez[1,3,4,*], Sarah Oh[2,3,*], Anthony Dickinson[5], Jose Arenas[1], & Emiliano Merlo[6]

[1] Department of Industrial Engineering, University of Chile, Santiago, Chile.

[2] Department of Psychology, University of California, Berkeley, CA, USA.

[3] Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA.

[4] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile.

[5] Department of Psychology and Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge, UK.

[6] School of Psychology, University of Sussex, Falmer, UK.

Author Note

Correspondence regarding this article should be sent to Omar D. Perez (omar.perez.r@uchile.cl) or Emiliano Merlo (e.merlo@sussex.ac.uk).

* Authors contributed equally to this work.

Abstract

Goal-directed actions are those performed with the expectation of producing a specific outcome and are therefore sensitive to changes in the outcome's current value without re-experiencing the action-outcome contingency. Disrupting the action-outcome association by a number of manipulations can reduce responding in humans, but if such interventions also affect the goal-directedness of the response remains unknown. Across two experiments, we trained participants to perform different actions for distinct outcomes. Participants then experienced either extinction (i.e., complete suspension of outcome delivery), or non-contingent training (i.e., outcomes were presented independently of responding). One outcome was then devalued before responding was tested in the absence of outcome presentations. An overall analysis of responding during this test revealed that outcome devaluation decreased responding to a degree that was not significantly reduced by either the extinction or non-contingent training. These results suggest that human goal-directed behavior is resistant to manipulations of the action-outcome contingency. We discuss these results in light of recent theories of goal-directed behavior.

Keywords: goal-directed, extinction, non-contingent learning, humans

The status of goal-directed actions after manipulations of the action-outcome contingency

Introduction

Humans, like other animals, rely on different systems to guide their actions. One is a goal-directed system, in which actions are selected based on their expected consequences; the other is a habitual system, in which behavior is driven largely by learned associations between situations and responses, often without reference to the current value of the outcome [1–9]. These systems are dissociable both computationally and neurally, and their interplay plays a role in many aspects of adaptive behavior.

A critical feature of goal-directed control is its sensitivity to changes in outcome value. In classic outcome devaluation paradigms, individuals learn to associate specific actions with rewarding outcomes. The value of one outcome is then reduced by satiating the animals on one of the outcomes, or aversive conditioning (when one of the outcomes is paired with gastric malaise) and behavior is then tested in the absence of outcome feedback. If an action is goal-directed, its selection will reflect the updated value of the outcome; if habitual, responding will persist regardless of devaluation. This dual-system approach has proven critical for understanding both healthy and impaired learning and decision-making, and has been used extensively in neuroscience and psychology [4,5,9,10].

Much of the evidence of these different types of behavioral control comes from experiments on rodents. In one particularly influential line of work, Prof. Rescorla[11] examined whether goal-directed control is maintained after the outcomes associated with an action are no longer delivered (i.e., under extinction). Surprisingly, he found that extinction did not weaken goal-directed control. Rats trained to associate different actions with different outcomes still adjusted their behavior according to the current value of the outcome, even when those actions no longer led to the rewarding outcomes and rats were

not performing the action by the end of extinction training. This finding suggested that extinction reduces responding by producing direct inhibition of the response, rather than erasing the original action–outcome association.

Subsequent work in animals has investigated other ways of degrading the action-outcome contingency. For instance, Crimmins et al.[12] examined whether presenting outcomes non-contingently—that is, independently of the subjects' behavior—would also preserve goal-directed control. They found that under certain conditions, such non-contingent training similarly failed to affect sensitivity to outcome devaluation, in spite of significantly reducing responding. The interpretation of their results, however, may depend on the specific contextual and procedural features of the task (see Discussion section).

Most dual-system theories assume the strength of the action-outcome association underlying goal-directed control should affect the degree to which changes in outcome value affect responding. Since goal-directed behavior is sensitive to the causal action-outcome association and the value of the outcome produced by the actions, any manipulation of the action-outcome strength should downweight the effect of changes in outcome value on responding[13,14]. The most prominent theory was postulated by Perez and Dickinson [2]. Using a goal-directed system that computes the action-outcome rate correlation across different working memory samples, coupled with a reinforcement learning habitual system, they were able to demonstrate that inhibition by the habit system can capture resistance to devaluation after extinction in spite of reductions in response rate, just as observed by Rescorla in animals. By contrast, their model anticipates that goal-directed behavior should not survive non-contingent training. No evidence has been provided for these two predictions in human subjects.

Given the theoretical importance of these findings, and the possibility that humans may employ more complex causal action–outcome representations and assign value to actions differently than other

species, we examined whether human goal-directed actions remain sensitive to devaluation after manipulations that disrupt the contingency between action and outcome. Specifically, we tested whether outcome devaluation continues to influence behavior after extinction (when outcomes are omitted) and after non-contingent training (when outcomes occur independently of responding). To do so, we adapted the classic devaluation paradigm to a free-operant setting in humans using a keypress task with token outcomes as rewards.

Methods

Participants and apparatus

We recruited participants from the Prolific online platform (www.prolific.com). The total number of participants for each experiment was 117 for the extinction experiment and 109 for the non-contingent experiment. As there were no previous experiments testing the impact of outcome devaluation on human free-operant responding after manipulations of the action-outcome contingency, we did not perform a power analysis before data collection. However, we aimed to reach 100 participants per experiment, which is considerably higher than similar designs on human action learning. Participants were required to be located in the US or UK, fluent in English, and were paid 8 GBP per hour for their participation. The study was approved by the University of Sussex Sciences & Technology Cross-Schools Research Ethics Committee (SCITEC) (approval ER/EM540/1).

Procedure

Our design was based on a paradigm originally developed by Rescorla [11]. The design of one of his experiments is illustrated in the extinction component of Table 1. Initially, he trained his hungry rats to perform four different instrumental responses with two of the responses (R1 and R3) being reinforced with one food outcome (O1), and the other two responses (R2 and R4) with a second food outcome

(O2). Having established responding, two of the responses (R1 and R2) were extinguished by omitting their respective outcomes. At issue was whether this extinction had weakened the R–O association mediating goal-directed control. However, in order to assess such control with an outcome revaluation test following extinction, responding had to be re-established using a third outcome (O3). In this way, the level of responding was comparable between extinguished and non-extinguished responses at the outset of the final test. In this test, Rescorla devalued O1 by conditioning a food aversion to it in the absence of the opportunity to perform any of the responses. Finally, the animals were given a choice between R1 and R2 and between R3 and R4 in the absence of any outcomes. Testing in absence of a contingent outcome is critical if responding is to reflect the extent to which devaluation impacts behavior via the R–O association, as any decrease in responding could otherwise be attributed to punishment by experiencing the aversiveness of the outcome after devaluation.

The fact that R3, originally trained with the now-devalued O1, was performed less on test than R4, trained with the still-valued O2, confirmed that the initial instrumental training established goal-directed control over responding. Surprisingly, however, the magnitude of the outcome devaluation effect revealed that the contrast between R1 and R2 was similar to that observed for R3 versus R4. Extinction of R1 and R2 by the total removal of their training outcomes had no effect on the extent to which the current value of these outcomes impacted performance. The theoretical implication is that extinction reduces responding by establishing a new R-no-outcome association, without decreasing the strength of the R–O association established during training.

Both experiments in this paper implemented the same experimental design outlined in Table 1 except for the fact that the instrumental contingency was degraded by omitting all outcomes in the extinction experiment and by presenting outcomes independent of responding in the non-contingent experiment. At the beginning of each experiment, participants were told that their goal was to press keys

to obtain coin outcomes of two different types, red and blue, and that their goal was to obtain as many coins as they could. They were also told that there was a third type of coin, green, which could also be earned in some blocks of the experiment, and that this coin was worth the same as the other two. Participants were also informed that the collected coins would be deposited in different fictitious piggy banks of limited size, one for each type of coin.

Table 1. **Design of experiments.** R and O denote responses and outcomes, respectively.

Training	Extinction or degradation	Retraining	Devaluation	Choice test
	Extinction experiment		Devalue O1	S1: R1 vs R2
S1:				
R1 → O1				
		R1 → O3		
R2 → O2	R1 → nothing	R2 → O3		
	R2 → nothing			S2: R3 vs R4
		R3 → O3		
		R4 → O4		
	Degradation experiment			
S2:				
R3 → O1				
R4 → O2	R1 → nothing / O1 free			
	R2 → nothing / O2 free			

Each experiment implemented the design illustrated in Table 1 in which two actions were exposed to the degraded contingency (R1 and R2), whereas the other two actions (R3 and R4) did not undergo any manipulation and functioned as within-subject controls. In each block of training, participants had the opportunity to press two keys, either the two top keys, "q" or "p", or the two bottom keys, "z" or "m". The location of the keys on the screen respected their relative position on the keyboard (e.g., "z" was shown on the bottom left side of the screen whereas "p" was shown on the top right location). For each participant, one of the top keys delivered red coins and the other one to blue coins. To avoid participants attributing the type of coin with the side of the response (which would have associated the two outcomes to different hands) the opposite assignment was programmed for the bottom keys. Therefore, actions performed with one of the hands always led to different outcomes across different blocks of training. For example, if for one participant the assignment for top keys was q → red and p → blue, then for the next block the assignment was z → blue and m → red.

The sequence of stages in the procedure is illustrated in Figure 1. Initially, participants underwent three 120-s blocks of concurrent training for each pair of actions. The starting pair of actions was randomized across participants and the blocks were presented sequentially once this first assignment was set per each participant. Each action delivered its outcome under a random-interval (RI) schedule, so that on average the appropriate coin for each response became available to be collected every T seconds. This was implemented by programming two independent clocks, each generating a reward with a probability of 1/T per second. The schedule for each action was capped so that a reward was guaranteed to be available for collection by a response if three times the scheduled inter-outcome interval T had elapsed. To minimize accidental reinforcement of switching between responses, we

introduced a changeover count: only the second response after a switch could produce a coin, even if the schedule on the new action had already made a reward available while the participant was engaged in the other response. The parameter T was set to 12 s for the extinction experiment and 15 s for the degradation experiment.

Once a total of three blocks of training for each pair of actions finished, participants were presented with a message informing them that the piggy banks were filling up (see “Devaluation” screenshot in Figure 1). This information was added to remind participants that their coins were being deposited and counted during the experiment and to familiarize them with the stimuli that would be presented at the end of the experiment after devaluation. After this screen, participants were presented with 30 coins located in random positions on the screen, 10 of each type, and asked to collect a total of 10 coins that would be deposited into their respective piggy banks. This free consumption phase (“Free coin collection” in Figure 1) was followed by the 15-min extinction or contingency degradation stage wherein two actions (either "q" and "p", or else "z" and "m") were presented and outcomes were omitted in the extinction experiment or followed by the non-contingent presentation of both outcomes each on an independent random time 15-s schedule in the non-contingent experiment. On this latter schedule, participants could achieve the same outcome rate irrespective of whether they responded or not, thereby generating a zero contingency between action and outcome[1].

After the manipulation phase, participants were given the opportunity to perform a 120-s block of retraining of all actions, which was similar to those in the training phase except for the fact that the outcome was the green coins for all responses. Participants were informed that these coins were worth the same as the blue and red ones, and that they would also be deposited into their respective piggy banks. In this manner, we ensured that participants were performing responses at comparable rates at the

[1] The use of a random time schedule to generate non-contingent outcomes differs from that employed by Crimmins et al (2022). They trained their rats on a ratio rather than interval schedule and therefore rendered the outcomes non-contingent by arranging that the outcome probability in the absence of a response was the same as that following a response.

end of training given that extinction or degradation should have reduced responding during the manipulation phase. This retraining phase was followed by devaluation of one of the outcomes. In this phase, participants were presented with a screen showing one of the piggy banks being closed so that no more coins of that type could be deposited in the piggy bank, effectively devaluing that outcome with respect to the others. To test if the devaluation phase succeeded in achieving this goal, we then presented participants with another consumption test where they were asked to collect 10 coins using mouse clicks in the same manner as the previous coin collection phase (see Figure 1). During this free collection phase, we expected that participants would collect less of the devalued coin than the two other still-valued coins.

Participants were then given the critical outcome devaluation test at the start of which they were told that they could respond to obtain coins in the same manner as in the training phase for 90 s, but that this time the outcomes obtained would be covered by a curtain (pseudoextinction). They were informed that, apart from that, nothing about the experiment had changed. The coins obtained would be still deposited into their respective piggy banks (provided they were not closed). Finally, participants underwent an additional reinforced test where the outcomes were again presented contingent to the actions. If human goal-directed behavior survives extinction[11] and non-contingent training[12], we expect similar devaluation effects in both experiments. We further expect the devaluation effect to be stronger in the reinforced test than in the pseudoextinction test.

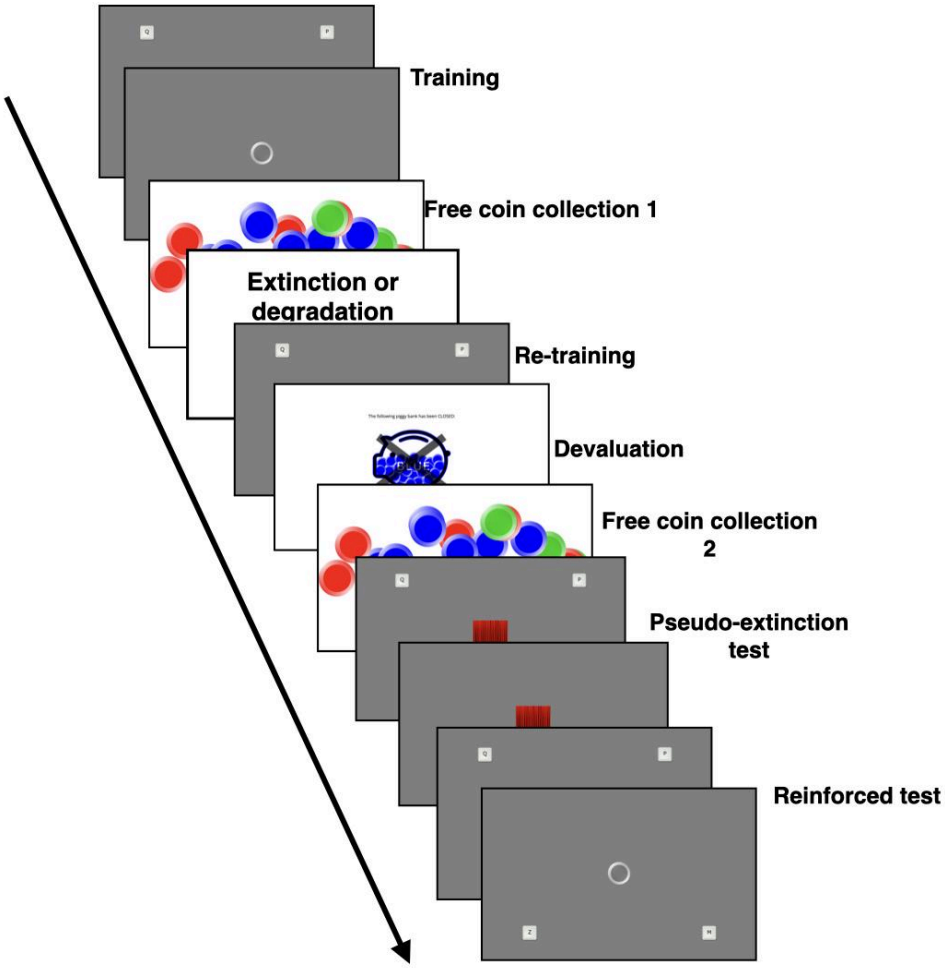

Figure 1. **Sequential presentation of stimuli in the task.** During training, participants were given 120 seconds of concurrent training to press either two top keys ("q" and "p") in one block of training, or two bottom keys ("z" and "m") in another block, followed by a re-training phase where they had the opportunity to press the two top or bottom keys concurrently. One of the two top keys led to red coins and the other to blue coins. The same applied to bottom keys.

Results

All our analyses were performed in the R programming language under the RStudio IDE[15]. We did not discard any participants based on their performance, unless they stopped responding in more than one block during the training blocks. No participants met this criterion.

A mixed-effects analysis was conducted using the *lmer* function from the lme4 package [16], with a random intercept and by-subject random slope for value (valued vs. devalued). The between-participant variable of experiment distinguished performance when the action–outcome contingency was degraded either by extinction or by non-contingent outcomes. Two within-participant variables were included: the contingency variable, contrasting degraded responses (R1 and R2) with control responses (R3 and R4), and the value variable, contrasting responses associated with the devalued outcome (R1 and R3) versus those associated with the still-valued outcome (R2 and R4). To assess the evidence in favor of the null, we computed likelihood ratios between models and Bayes factors using the bayesfactor_models from the bayestestR package [17].

Figure 2A shows the mean response rates during the devaluation choice test for each condition. In both experiments, participants responded less for actions associated with the devalued outcome than for those associated with a still-valued outcome. However, neither the numerical means nor the statistical interactions indicated that the contingency manipulation reduced the magnitude of this devaluation effect. The analysis revealed a robust main effect of value ($b = 29.5$, $SE = 7.19$, $p < .001$), indicating that participants responded more for valued than for devalued outcomes. However, there was no main effect of contingency ($b = 3.81$, $SE = 3.84$, $p = .340$), and no significant interaction between value and contingency ($b = –5.69$, $SE = 5.43$, $p = .313$). The interaction between value and experiment was also not significant ($b = –18.3$, $SE = 9.99$, $p = .068$), nor did the three-way interaction among value, contingency, and experiment ($b = 7.06$, $SE = 7.83$, $p = .368$). These results indicate that the magnitude of the outcome devaluation effect did not reliably differ between extinction and non-contingent training conditions.

To further assess whether the value effect differed across experiments, we formally compared a full model including the three-way interaction among value, contingency, and experiment to a reduced

model excluding this interaction. A likelihood ratio test comparing the models yielded no significant improvement in fit for the full model, $\chi^2(3) = 3.43$, $p = .330$, and the Bayes factor strongly favored the reduced model over the full model (BF = 4890, BIC approximation), providing strong evidence for model parsimony[18]. This reinforces the conclusion that the magnitude of the devaluation effect did not differ reliably between extinction and non-contingent training conditions.

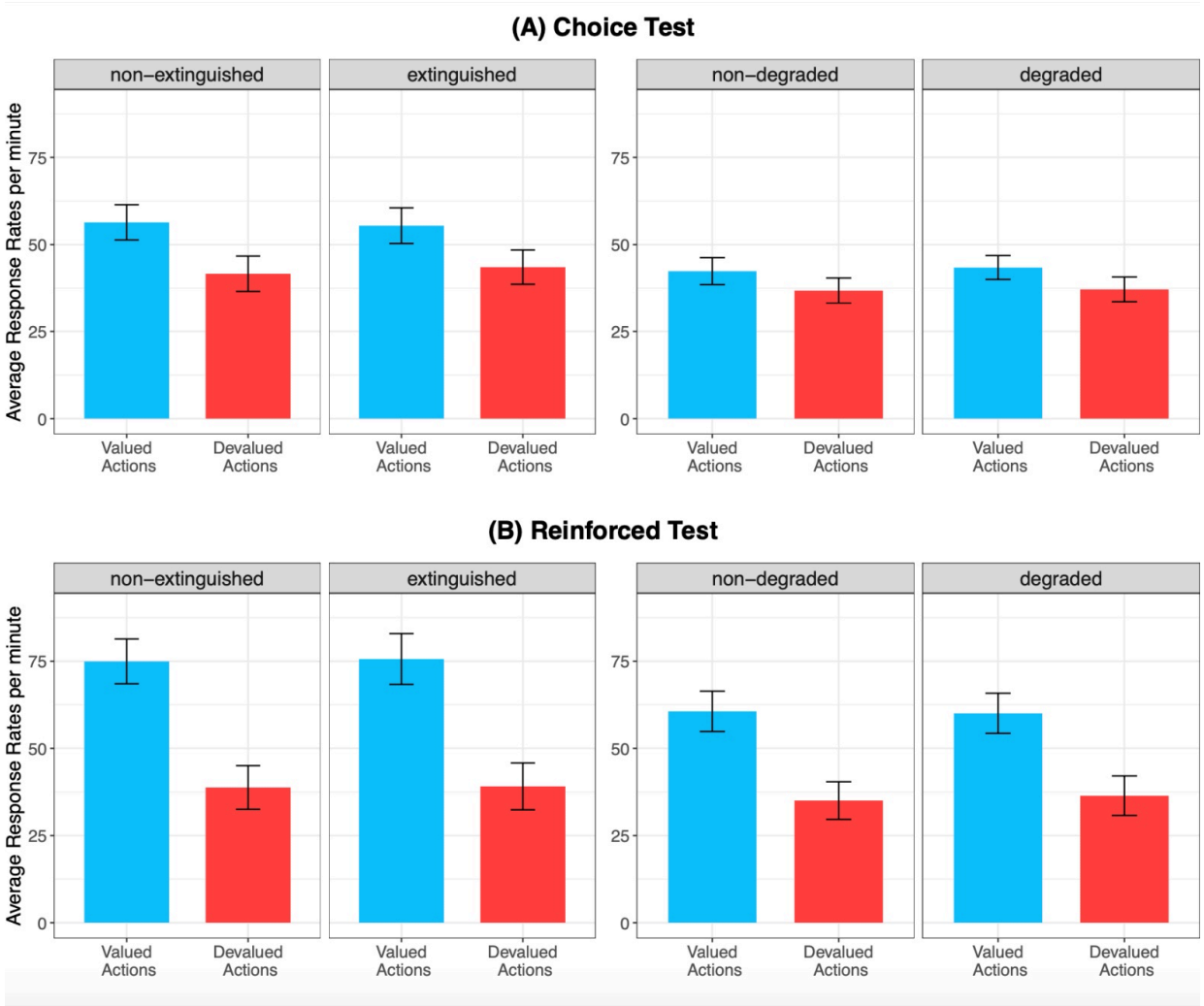

Figure 2. **Response rates in the tests for extinction and degradation experiments. A)** Average responses rates during the pseudo-extinction choice tests. **B)** Average response rates during the reinforced tests. Error bars represent 95% within-subject confidence intervals[19].

The panels of Figure 2B show the response rates during the reinforced test, in which participants received feedback on their actions (the outcomes were reintroduced contingent on responding, as during training). The same value-based pattern was observed: participants responded more for valued than for devalued outcomes. The value effect was numerically enhanced relative to the choice test, as indicated by a significant main effect of test phase ($b = 42.8$, $SE = 7.56$, $p < .001$) and a value × phase interaction ($b = 42.8$, $SE = 7.56$, $p < .001$), confirming that the devaluation effect was stronger in the reinforced than in the extinction-like test. However, this effect did not interact significantly with experiment ($b = 2.15$, $SE = 7.43$, $p = .775$), suggesting that the pattern was similar across extinction and degradation manipulations. Likewise, there was no reliable three-way interaction among value, phase, and experiment ($b = –11.7$, $SE = 14.9$, $p = .433$), indicating that the enhancement of the devaluation effect during the reinforced test generalized across both experimental conditions.

Figure 3A shows the final mean response rates during the last 2 min of each phase of the extinction and degradation experiments. An experiment × value × contingency analysis of these final training blocks revealed no significant main effects or interactions. This indicates that performance prior to the manipulation phase was comparable across conditions.

Figure 3B shows response rates across the contingency manipulation phase, separated by experiment. As expected, participants progressively reduced their response rate for the degraded actions over time. However, degraded response rates remained higher under non-contingent training than under extinction by the end of the manipulation phase, a difference that is also usually found in the animal literature. This description was supported by a significant Experiment × Time-block interaction ($b = -.252$, $SE = .037$, $p <.001$), confirming that response suppression during the manipulation was stronger in the extinction condition.

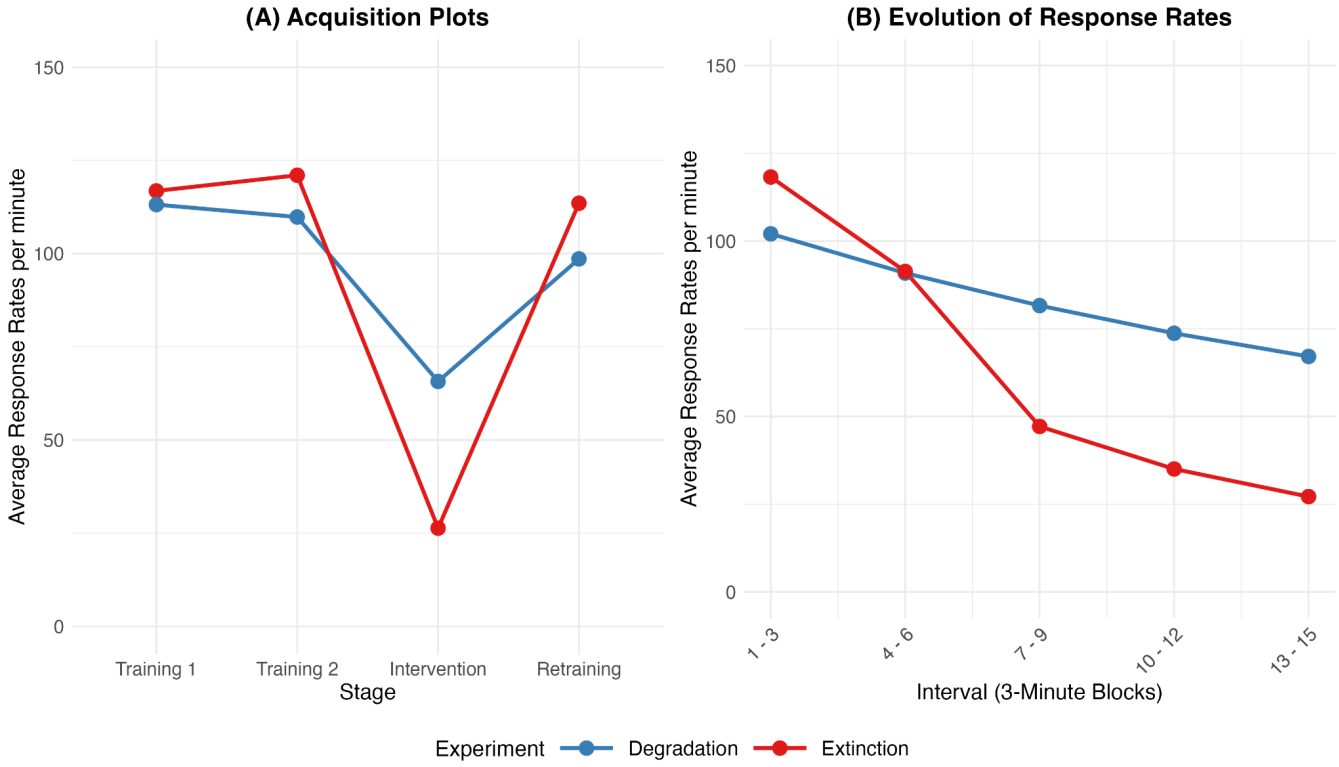

Figure 3. **Response rates for training and during the intervention phase. A)** Response rates during training, intervention and retraining phases. Means are calculated for the last 2 minutes of each phase. **B)** Response rates during the intervention phases. Means are calculated for each of the 5 3-min blocks.

To explore whether individual differences in behavioral suppression during the manipulation phase predicted value sensitivity, we included the response suppression score (response rate in the last 2 min of the intervention phase minus response rate in the last 2 min of the last block of training) in the model and tested its interaction with outcome value. This analysis revealed a significant value × suppression interaction ($b = -0.171$, $SE = 0.076$, $p = .025$), indicating that participants who showed greater reductions in responding during the manipulation phase also showed a stronger devaluation effect. However, this effect did not interact reliably with experiment or contingency conditions. Thus, while the degree of suppression covaried with goal-directed performance across individuals, this relationship was consistent across both extinction and non-contingent training.

Discussion

In two experiments, we tested in humans how manipulations of action-outcome contingency affect sensitivity to devaluation of the outcomes produced by the actions. In the extinction experiment, both extinguished and non-extinguished actions were devaluation sensitive, which is consistent with the results obtained by Rescorla[11] in rodents. Similarly, the second experiment replicated with human participants the Crimmins et al.[12] finding that the degradation of the instrumental contingency by training with non-contingent outcomes can also have no impact on the rodent outcome devaluation effect, at least when the actions involve responses on different manipulanda (see below).

The theoretical interpretation of these findings is complex. The theory of goal-directed learning offered by Perez and Dickinson[2] is unique in providing an account of why such learning is resistant to extinction. Under this theory, it is the mean local correlation between response and outcome rates that determines the strength of goal-directed control. The local correlation is estimated across time samples in a working memory which records the number of responses and outcomes in each sample. Every time the memory recycles by starting a new sample and dropping an old one, the local correlation in the memory immediately prior to the recycle is estimated and contributes to a running average of the local rate correlation experienced by the agent.

The primary motivation for developing this theory was the fact that ratio schedules of reinforcement, under which this correlation is positive, tend to produce higher response rates and devaluation sensitivity than interval schedules, which generate relatively low correlations. However, an incidental prediction of the theory is that extinction should have only a limited impact on goal-directed control. The rate correlation can only be computed when both responses and outcomes are represented in memory and, of necessity, the computation is halted when the memory is cleared of representations of at least one of the event types, either the response or the outcome. Given that this system only computes

the local rate correlation, this suspension should happen relatively early in extinction with loss of any outcome representation and thereafter the strength of goal-directed control will reflect the mean running average rate correlation established earlier during training. Perez and Dickinson verified this incidental prediction of the theoretical model by formal simulation[2].

More problematic for Perez and Dickinson's theory is that response reduction by non-contingent outcomes had no detectable effect on goal-directed control, as demonstrated by the Crimmins et al.[12] rodent experiments and our non-contingent human experiment. In fact, Perez and Dickinson explicitly offered the prediction that non-contingent training, in contrast to extinction, would reduce goal-directed control as a novel prediction of their theory and therefore these results may well challenge rate correlation theory. However, as Crimmins et al.[12] point out, classic two-process theory[20] has long argued that an instrumental contingency often arranges not only an action-outcome relationship but also a Pavlovian stimulus-outcome association that endows the stimulus with the capacity to modulate instrumental performance. As a result, what appears to be an instrumental outcome revaluation effect may not be mediated by action-outcome learning as required for goal-directed control but rather by Pavlovian stimulus-outcome learning. Although Jonkman et al.[21] provided evidence that Pavlovian contextual conditioning can contribute to apparent goal-directed control, Crimmins et al.[12] focus their analysis on the role of their manipulanda, right- and left-hand levers.

This focus reflected the fact that the procedures in the experiments in which they found that non-contingent training had no effect on the outcome devaluation effect were ones in which each response was trained in separate sessions with only the appropriate lever present so that presence of a specific lever predicted the type of outcome. Evidence that the presence of a lever can act as a Pavlovian

---

[2]The reader may wonder why responding is reduced by extinction if goal-directed strength is maintained. However, Perez and Dickinson (2020) embedded the goal-directed system within a dual-system theory that includes a habitual system. Within this dual-system account, extinction is primarily due to the development of response inhibition by the habitual system, which accords with the explanation offered by Rescorla[11].

stimulus comes from a study by Amaya, Stott, and Smith[22] in which the insertion of a lever signaled the delivery of a food outcome. Although there was no instrumental contingency in that the rats did not have to press the lever to receive the outcome, they regularly learned to approach the lever and contact it and thus often press it. Critically, this Pavlovian lever pressing was mediated by a lever stimulus-outcome association in that conditioning an aversion to the food outcome yielded an outcome devaluation effect on test.

Crimmins et al.[12] removed this confound by using a single lever which when pressed to the left yielded one outcome with the other outcome yielded on right presses. With this bi-directional lever that equally predicted the two outcomes, the results were those anticipated by rate correlation theory, namely that non-contingent training reduced the outcome devaluation effect and thus evidence for goal-directed control. As our online participants did not have access to a bidirectional manipulandum, we ensured that the stimuli accompanying each pair of responses (R1/R2 and R3/R4) had the same Pavlovian associations with both outcomes by training the two responses concurrently so that all stimuli were equally associated with both outcomes. This procedure ensured that it is unlikely that the outcome devaluation effect was mediated by Pavlovian associations with the outcome.

Although both extinction and non-contingent training significantly reduced responding, it is notable that extinction produced a greater reduction[23]. However, it is also notable that there was still considerable responding after both contingency manipulations, but we think that it is unlikely that greater reductions in responding would have decreased the outcome devaluation effects. Further analyses that included the magnitude of response reduction during extinction and non-contingent training as variables revealed that the size of the outcome devaluation effect increased with greater response reductions. This effect probably reflects individual differences in the cognitive resources applied to the task—the greater the resource allocation to encoding the relative contingencies, the more

likely it is that a participant will have reduced responding under extinction and non-contingent training and learned the response–outcome associations and outcome values necessary for the devaluation effect. However, the correlation between suppression and contingency knowledge was not significant ($r = .01$, $p = .667$), suggesting that these variables reflect different processes. When both were included in the model, each independently predicted sensitivity to outcome devaluation: suppression ($b = –0.171$, $SE = 0.076$, $p = .025$) and contingency score ($b = –2.32$, $SE = 0.83$, $p = .005$). This suggests that while suppression may capture attentional engagement or strategic adaptation during the manipulation phase, contingency scores appear to reflect explicit knowledge of instrumental structure. However, they both appear to contribute independently to goal-directed control.

Our paradigm is the first one in humans including a reinforced test at the end of training (but see Gera et al. [24] for an exception in an fMRI experiment). Previous studies in humans have ignored this fundamental test in assessing the effect of value on responding after manipulations aimed to prompt habit formation and relied either on consumption tests similar to our coin collection procedure or specific satiety [25–27]. Although this test is necessary to show the effectiveness of devaluation, it is not sufficient: what is more critical is showing that devaluation is effective in the same training context and contingent on the actions performed during training. Our data show that the devaluation method was effective in doing so by revealing a stronger effect of outcome value in the reinforced choice test than in the pseudoextinction choice test.

In conclusion, our findings show that human goal-directed behavior is resistant to disruptions in action–outcome contingency, just as for manipulations of training extension [27–30]. While extinction and non-contingent training reduce response rates, they do not eliminate value-based control. This challenges current theories and points to the robustness of goal-directed systems in humans. It seems

that, under many circumstances, previous expectations of producing a valued outcome are remarkably difficult to erase.

Author note

Omar David Perez is supported by Instituto de Sistemas Complejos de Ingeniería (ISCI) ANID PIA/PUENTE AFB220002, ANID-SIA 85220023 and ANID-FONDECYT 1231027. Emiliano Merlo is supported by the University of Sussex School of Psychology start-up funds.

Author contributions

ODP, SO, AD and EM designed the study. SO programmed the tasks. ODP and JA analyzed the data. ODP and AD drafted the manuscript, with critical editions by all authors.

References

1. Daw, N. D. & O'Doherty, J. P. Multiple Systems for Value Learning. in *Neuroeconomics: Decision Making and the Brain: Second Edition* 393–410 (2013). doi:10.1016/B978-0-12-416008-8.00021-8.
2. Perez, O. D. & Dickinson, A. A Theory of Actions and Habits: The Interaction of Rate Correlation and Contiguity Systems in Free-Operant Behavior. *Psychol. Rev.* (2020) doi:10.1037/rev0000201.
3. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
4. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
5. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
6. Henriquez-Jara, B., Guevara, C. A., Munizaga, M. & Perez, O. D. Habits and the subexploration of better transportation options: A dual-system approach. *Travel Behav. Soc.* **38**, 100877 (2025).
7. Everitt, B. J. & Robbins, T. W. Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annu. Rev. Psychol.* **67**, 150807174122003 (2016).
8. Everitt, B. J. & Robbins, T. W. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* **8**, 1481–1489 (2005).
9. Yin, H. H. & Knowlton, B. J. The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* **7**, 464–476 (2006).
10. Tricomi, E., Balleine, B. W. & O'Doherty, J. P. A specific role for posterior dorsolateral striatum in human habit learning. *Eur. J. Neurosci.* **29**, 2225–2232 (2009).
11. Rescorla, R. A. Preservation of response-outcome associations through extinction. *Anim. Learn. Behav.* **21**, 238–245 (1993).
12. Crimmins, B. *et al.* Response-independent outcome presentations weaken the instrumental response-outcome association. *J. Exp. Psychol. Anim. Learn. Cogn.* **48**, 396 (2022).

13. Urcelay, G. P. & Jonkman, S. Delayed rewards facilitate habit formation. *J. Exp. Psychol. Anim. Learn. Cogn.* **45**, 413 (2019).
14. Perez, O. D. & Urcelay, G. P. Delayed rewards weaken human goal-directed actions. *Npj Sci. Learn.* (2025).
15. RStudio Team. RStudio: Integrated Development Environment for R. (2020).
16. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
17. Makowski, D., Ben-Shachar, M. S. & Lüdecke, D. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *J. Open Source Softw.* **4**, 1541 (2019).
18. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
19. Morey, R. D. Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutor. Quant. Methods Psychol.* **4**, 61–64 (2008).
20. Rescorla, R. A. & Solomon, R. L. Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychol. Rev.* **74**, 151 (1967).
21. Jonkman, S., Kosaki, Y., Everitt, B. J. & Dickinson, A. The role of contextual conditioning in the effect of reinforcer devaluation on instrumental performance by rats. *Behav. Processes* **83**, 276–281 (2010).
22. Amaya, K. A., Stott, J. J. & Smith, K. S. Sign-tracking behavior is sensitive to outcome devaluation in a devaluation context-dependent manner: implications for analyzing habitual behavior. *Learn. Mem. Cold Spring Harb. N* **27**, 136–149 (2020).
23. Rescorla, R. A. & Skucy, J. C. Effect of response-independent reinforcers during extinction. *J. Comp. Physiol. Psychol.* **67**, 381 (1969).
24. Gera, R. *et al.* Characterizing habit learning in the human brain at the individual and group levels: A

multi-modal MRI study. *NeuroImage* **272**, 120002 (2023).

25. Gillan, C. M. M., Otto, A. R., Phelps, E. A. & Daw, N. D. Model-based learning protects against forming habits. *Cogn. Affect. Behav. Neurosci.* **15**, 523–536 (2015).
26. Pool, E. R., Pauli, W. M., Kress, C. S. & O'Doherty, J. Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans. *Nat. Hum. Behav.* **3**, (2019).
27. Pool, E. R. *et al.* Determining the effects of training duration on the behavioral expression of habitual control in humans: a multilaboratory investigation. *Learn. Mem.* **29**, 16–28 (2022).
28. Gera, R., Barak, S. & Schonberg, T. A novel free-operant framework enables experimental habit induction in humans. *Behav. Res. Methods* **56**, 3937–3958 (2023).
29. de Wit, S. *et al.* Shifting the balance between goals and habits: Five failures in experimental habit induction. *J. Exp. Psychol. Gen.* **147**, 1043 (2018).
30. Garr, E., Padovan-Hernandez, Y., Janak, P. H. & Delamater, A. R. Maintained goal-directed control with overtraining on ratio schedules. *Learn. Mem.* **28**, 435–439 (2021).