# INVESTIGATING THE ROLE OF REPRESENTATION SWITCHING COSTS IN GOAL PERSISTENCE BIAS

**Gaia Molinaro**[*]
Department of Psychology, Neuroscience Department
University of California, Berkeley
Berkeley, CA 94720
gaiamolinaro@berkeley.edu

**Aly Lidayan**[*]
Department of EECS
University of California, Berkeley
Berkeley, CA 94720
dayan@berkeley.edu

**Anne G. E. Collins**
Department of Psychology, Neuroscience Department
University of California, Berkeley
Berkeley, CA 94720
annecollins@berkeley.edu

[*]Co-first authors

## ABSTRACT

Goal pursuit profoundly shapes human cognition, typically benefiting learning and decision-making by focusing information processing. However, goal-dependent processing occasionally leads to seemingly maladaptive behavior, such as a bias toward *goal persistence* – the tendency to continue pursuing the current goal even when suboptimal. While various explanations have been proposed for such goal persistence, the underlying cognitive mechanisms remain unclear. We hypothesize that one key factor is the computational cost of switching between different internal representations of external stimuli that is induced by changes in the active goal. Humans' bias towards persisting with suboptimal goals could thus be resource-rational concerning their cognitive capacities. To test this, we developed a task where participants chose between competing goals, with rewards structured to encourage goal switching. Goals were organized in pairs requiring the same rules for action selection. In our initial experiment (N = 67), participants showed both a bias towards goal persistence and a preference for switching between goals sharing the same rule. This preference correlated with individual differences in cognitive flexibility, i.e., a stronger preference was positively correlated with greater switch costs (i.e., increased reaction times) following different-rule switches compared to same-rule switches. Our preliminary results are consistent with our hypothesis that representation switch costs may play a role in goal persistence biases. In continuing the project, we will recruit additional participants on modified versions of the task. We predict that 1) we will replicate previous findings showing goal persistence; 2) goal switching will be biased toward same-rule goals; 3) this bias will correlate negatively with participants' cognitive flexibility; and 4) goal persistence biases will increase when alternative goals follow different rules compared to the current goal.

## 1 INTRODUCTION

Goals have a profound effect on human cognition. Setting a goal induces a shift in mindset, inducing an attentional shift towards the active goal and its associated information (Gollwitzer, 1990; Holton et al., 2024; O'Reilly, 2020; Sepulveda et al., 2020) and affecting internal representations of states, actions, and rewards (Molinaro & Collins, 2023c). This is often beneficial. For instance, warping state spaces in a goal-consistent manner increases the efficiency of goal-directed choices (Castegnetti et al., 2021), and coding rewards in a goal-congruent fashion benefits learning and decision-making in the service of the active goal (Frömer et al., 2019; McDougle et al., 2022; Molinaro & Collins, 2023b). However, goal-dependent biases may also cause irrational or suboptimal decisions. For instance, goal-dependent rewards can lead to differences in the valuation of the different options despite their equivalent expected value (Molinaro & Collins, 2023a). Moreover, fixating on the current goal can result in a failure to disengage from unfavorable objectives that would best be abandoned – a well-established phenomenon in human and other animals' behavior known as the "sunk cost fallacy" (Holton et al., 2024; Sweis et al., 2018).

Understanding why such apparently maladaptive goal persistence occurs is crucial for both basic research and practical applications. For instance, more accurate models of human (ir)rationality could foster cooperation between human and artificial agents. While current artificial systems model human behavior according to normative principles (Liu et al., 2025), building machines that understand people requires instilling knowledge about systematic human biases into them (Bobu et al., 2020; Howes et al., 2023; Jacob et al., 2024). Accurately accounting for human preferences in goal pursuit could allow for more helpful artificial assistants.

Several accounts of goal persistence have been offered to date. Under one explanation, sunk cost biases arise from a desire to avoid appearing wasteful to others (Arkes & Blumer, 1985). However, evidence for the phenomenon has also been found under non-social experimental conditions. A recent study found attentional shifts to correlate with goal persistence biases, but did not discuss the underlying source of such attentional switches (Holton et al., 2024). Here, we propose costs in task representation switches as one reason people keep pursuing a goal even when it is maladaptive to persist. In this context, a representation refers to a specific configuration of the cognitive system required to perform a task (Schneider & Logan, 2007). We ground our hypothesis on two well-established phenomena. First, there is broad evidence that goal setting induces a particular neural state in which internal representations of states are warped to align with the current objective (Castegnetti et al., 2021; Molinaro & Collins, 2023c; O'Reilly, 2020). Second, decades of cognitive control literature show that reconfiguring such a state is computationally and energetically expensive (Grahek et al., 2023; Gu et al., 2017), and behaviorally costly (Monsell & Mizon, 2006).

To test our proposal, we developed an experimental paradigm inspired by Holton et al. (2024) in which participants selected between competing goals on each trial. We designed the task so that in some trials, persisting with the current goal would yield a lower expected reward rate than switching to an alternative goal. Each goal required participants to select actions based on specific stimulus features (different "rules"), and goals were organized in pairs that shared the same rule for action selection. We predicted that when selecting new goals, participants would prefer goals that shared the same rule as their last goal, to avoid the cost of switching representations. In addition to replicating past results on goal persistence Holton et al. (2024), our preliminary data (N = 67) are consistent with this prediction, showing that participants indeed preferred same-rule goals. Additionally, we found initial evidence that these same-rule preferences correlate negatively with cognitive flexibility, suggesting that goal-switching preferences might be adaptive to individual differences in cognitive abilities, consistent with resource-rational accounts of goal pursuit (Prystawski et al., 2022). In future iterations, we will manipulate the availability of same-rule goals to test our prediction that participants showing stronger representation switching costs also exhibit more goal perseveration.

## 2 EXPERIMENTAL PROCEDURE

In this task (Figure 1), participants collected gems to complete different colored necklaces. On each trial, participants first selected one of four possible necklaces (goals), each requiring a specific number of gems to complete. Participants were then presented with two keys, each characterized by a shape (heart or clover) and a pattern (dotted or striped). Selecting the correct key yielded gems matching the chosen necklace's color, incrementing progress toward completing the goal. To induce participants to occasionally switch away from initiated goals, the offer quantity (quantity of gems obtained from correct choices) for the current in-progress necklace could suddenly decrease, and then remain at the depleted level until necklace completion. Switching to a new necklace restored the depleted offer quantity but reset progress on the previously selected necklace. Baseline offer quantities were equal across gem colors and stable within each of the three blocks (80 trials each), while depleted offer quantities could vary on each event, but were always lower than baseline amounts. The threshold of goal progress at which offers were depleted also varied on each occasion. Across the three blocks, baseline offer quantities were either 3, 4, or 5, for necklaces of size 15, 20,

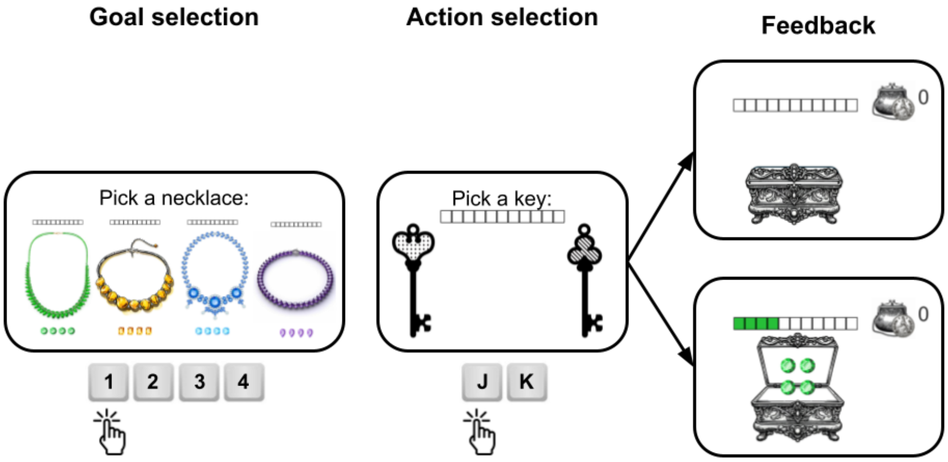

Figure 1: Task structure. On each trial, participants chose a goal (necklace) using number keys 1-4. Next, they chose an action (key) by pressing “J” or “K” on their keyboard. The total number of gems needed to complete each necklace and the number of gems available for each necklace on the current trial are shown. Keys are characterized by a shape and pattern. Finally, participants are shown either an empty treasure box (if they press the incorrect key) or the gems obtained (correct key) and updated progress toward completing the chosen necklace, along with the total rewards earned thus far in the game.

and 20, respectively, with block order pseudo-randomized across participants. Following depletion, offer quantities for the goal currently being pursued could drop to 1 or 2. Depletion occurred during roughly 30% of goal pursuits, with the depletion beginning at varying degrees of goal completion, ranging from 20%-60%. This range of offers, sizes, drop amounts, and goal completion stages were manipulated to allow us to estimate which factors contribute to participants’ likelihood of abandoning a goal. Upon collecting a number of gems equal to or higher than the amount required to complete a necklace, participants received 100 points, after which progress toward the chosen necklace was reset to 0 and its offer quantity was reset to the baseline.

Only one key feature was relevant for finding gems of a particular color. Of the four necklaces, two required selecting keys based on pattern and two based on shape. The correct pattern or shape (“rule”) was matched for the two keys sharing a key feature (for example, green and yellow required selecting the stripe-patterned key; blue and purple the clover-shaped key). Therefore, the four necklaces were created in two pairs with matching action selection rules.

To avoid potential order effects in learning the correct rule for each necklace, participants underwent extensive training prior to the main task. During training, participants completed short (12-trial) practice blocks for each necklace separately, and one practice block where they were forced to select specific necklaces, in a random order. Participants repeated each practice block until they reached 80% accuracy, for a maximum of 3 times. This procedure ensured that, by the end of practice, each participant had learned the rules associated with each necklace.

Participants were compensated with $12 for approximately one hour of their time. No additional performance bonuses were disbursed.

## 3 RESULTS

Participants were selected from Prolific (`https://www.prolific.com/`). We collected data from a total of 78 participants. We excluded 6 for achieving less than 80% accuracy in the main task key selection, 3 for failing to select a necklace on more than 5 trials, and 2 for failing to select a key more than 10 times, all indicating poor task compliance. Therefore, data from 67 participants were analyzed.

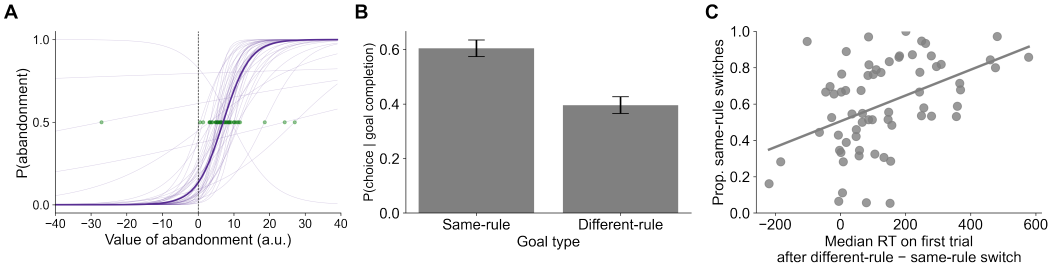

Figure 2: Behavioral results. A) Compared to an optimal agent, participants were more biased in persisting with the current goal. Transparent lines show the model fits of individual participants, the bolded line shows the mean model fit across all participants. The green dots show the value at which each subject was indifferent towards abandonment or persistence; this was significantly above 0, which is the optimal abandonment point. B) When choosing their next goal, participants were, on average, more likely to select a goal with the same rule as the previous one. Error bars indicate the SEM. C) Choice reaction time costs of switching to different-rule goals correlated with preferences for selecting same-rule goals.

### 3.1 Goal persistence bias

We first replicated the findings of Holton et al. (2024) that humans display a bias towards goal persistence. Following Holton et al. (2024), we modeled the value of goal abandonment as the estimated decrease in the number of steps until a goal is completed (and reward is obtained) if an alternative goal is picked over the current goal, $\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}$. Holton et al. (2024) used a stochastic tree-search model to estimate this. We adapted this approach to our task, which enabled two simplifications. First, we only analyze choices made for trials where the goal offer quantity is depleted, since alternative goals offer no advantage compared to the current goal before depletion. Second, we leveraged the fact that, in our task, if it is optimal to persist with a goal following offer depletion, this will remain the case until goal completion, since the depleted offer quantity cannot further decrease and the goal progress can only increase. Thus, we estimated $\hat{T}_{\text{current}}$ as the number of gems needed to complete the necklace divided by the depleted offer quantity (i.e., the number of steps to complete the current goal assuming no future switches and perfect choice accuracy). Since the four necklaces share a constant baseline offer quantity and size throughout the block, we estimated $\hat{T}_{\text{alt}}$ as the expected number of steps needed to complete a goal from scratch in the current block assuming perfect choice accuracy.

We found that, on average, participants abandoned their goals in 18% of the trials following gem depletion. In those trials, switches were more likely the more drastic the depletion, as evidenced by a significant effect of offer difference on average switch probability ($\beta$ = 0.19 ± 0.06, p = 0.002 from a linear mixed effects model with a fixed effect of offer difference and a random effect of participant identity).

To further quantify goal persistence biases, we again follow Holton et al. (2024) to model the subjective value of goal abandonment and probability of abandonment as follows:

$$SV_{\text{abandon}} = \beta_0 + \beta_1(\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}); \qquad P_{\text{abandon}} = \frac{1}{1 + \mathrm{e}^{-SV_{\text{abandon}}}}. \tag{1}$$

$$SV'_{\text{abandon}} = \beta_0 + \beta_1(\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}) + \beta_2 C_{\text{switch}}; \qquad P_{\text{abandon}} = \frac{1}{1 + \mathrm{e}^{-SV'_{\text{abandon}}}}. \tag{2}$$

We fit this model in a mixed effects logistic regression analysis predicting abandonment choices with random intercepts and slopes for each participant (mean cross-validation accuracy = 71%). As expected, we find that nearly all subjects showed a bias towards goal persistence (Figure 2A).

### 3.2 Representation switch avoidance

Because it does not impact the relevant feature for correct action selection, switching to a same-rule goal involves a smaller change in internal representations compared to switching to a different-rule goal. Therefore, we expected participants to switch to same-rule goals more often than different-rule goals. When choosing the next goal after goal completion, the majority of participants preferred avoiding rule changes, on average selecting same-rule goals 60% of the time, which was higher than chance (i.e., 50% t(66) = 3.39, p = 0.001; Figure 2B).

### 3.3 Resource rationality in goal selection

As expected, we observed a significant reaction time switch cost (Rogers & Monsell, 1995) when participants selected a different- vs. same-rule goal, measured as the difference between median choice reaction times (RTs) following goal completion or abandonment in same-rule trials (including same-goal trials; M = 690.44 ± 17.74) vs. different-rule trials (M = 823.79 ± 19.96, t(65) = -6.77, $p < 0.001$). We predicted that individual differences in participants' ability to efficiently transition between different goal-dependent representations might impact their goal selection. Indeed, we found that the greater the switch cost, the higher the proportion of same-rule switches ($\rho(65)$ = 0.46, $p < 0.001$; Figure 2C). One possible explanation for this pattern of results is that participants adjusted goal selection to their cognitive capacities, in line with a resource-rational account of goal pursuit (Prystawski et al., 2022).

## 4 Future directions

Having found initial support for our proposal that representation switch costs affect goal selection, we plan to run modified versions of the current task that will allow us to test specific hypotheses even more directly. We expect to confirm our initial behavioral results that participants 1) will display a goal persistence bias as demonstrated in previous studies; 2) will prefer switching to same-rule compared to different-rule goals; 3) will have a stronger preference for same-rule goals when experiencing greater goal-directed choice costs following different-rule switches. Moreover, we will set up alternative versions of the task to manipulate the availability of same-rule goals, enabling us to test the additional prediction of 4) a positive correlation between goal persistence bias and representation switching costs. In addition, we intend to develop a computational model sensitive to representation switch costs to capture participants' trial-by-trial goal selection. Finally, future experiments may focus on settings that enable specific distinctions between representation switch costs related to goal-conditioned policies vs. state representations.

## 5 Conclusion

Our initial findings suggest that the computational costs of switching goal representations may influence goal selection and contribute to goal persistence behavior. We also find preliminary evidence supporting a resource-rational account of goal pursuit, where representation switching costs lead to more conservative goal selection as a function of individual cognitive abilities. These results expand our understanding of why humans sometimes persist with suboptimal goals beyond social and attentional explanations. Future work with larger samples and a computational model that formalizes behavioral findings will be needed to validate our initial results and explore their broader implications for theories of goal-directed behavior.

### Acknowledgments

This work was funded by a 2023 Grant Award from the UC Berkeley Institute of Cognitive and Brain Sciences awarded to GM and AL. AL is also supported by the Fannie and John Hertz Foundation; GM and AGEC are also supported by NSF BCS 2336466.

## References

Hal R Arkes and Catherine Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140, 1985.

Andreea Bobu, Dexter RR Scobee, Jaime F Fisac, S Shankar Sastry, and Anca D Dragan. Less is more: Rethinking probabilistic models of human behavior. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pp. 429–437, 2020.

G. Castegnetti, M. Zurita, and B. De Martino. How usefulness shapes neural representations during goal-directed behavior. *Sci. Adv.*, 7(15):eabd5363, April 2021.

Romy Frömer, Carolyn K. Dean Wolf, and Amitai Shenhav. Goal congruency dominates reward value in accounting for behavioral and neural correlates of value-based decision-making. *Nat. Commun.*, 10(1):4926, October 2019.

Peter M Gollwitzer. Action phases and mind-sets. *The Handbook of Motivation and Cognition: Foundations of Social Behavior*, 2, 1990.

Ivan Grahek, Xiamin Leng, Sebastian Musslick, and Amitai Shenhav. Control adjustment costs limit goal flexibility: Empirical evidence and a theoretical account. *bioRxiv*, pp. 2023–08, 2023.

Shi Gu, Richard F Betzel, Marcelo G Mattar, Matthew Cieslak, Philip R Delio, Scott T Grafton, Fabio Pasqualetti, and Danielle S Bassett. Optimal trajectories of brain state transitions. *NeuroImage*, 148:305–317, 2017.

Eleanor Holton, Jan Grohn, Harry Ward, Sanjay G Manohar, Jill X O'Reilly, and Nils Kolling. Goal commitment is supported by vmPFC through selective attention. *Nature Human Behaviour*, pp. 1–15, 2024.

Andrew Howes, Jussi PP Jokinen, and Antti Oulasvirta. Towards machines that understand people. *AI Magazine*, 44(3):312–327, 2023.

Athul Jacob, Abhishek Gupta, and Jacob Andreas. Modeling boundedly rational agents with latent inference budgets. In *International Conference on Learning Representations*, 2024.

Ryan Liu, Jiayi Geng, Joshua C Peterson, Ilia Sucholutsky, and Thomas L Griffiths. Large language models assume people are more rational than we really are. In *ICLR*, 2025.

Samuel D McDougle, Ian C Ballard, Beth Baribault, Sonia J Bishop, and Anne G E Collins. Executive function assigns value to novel goal-congruent outcomes. *Cereb. Cortex*, 32(1):231–247, January 2022.

Gaia Molinaro and Anne G. E. Collins. Intrinsic rewards explain context-sensitive valuation in reinforcement learning. *PLOS Biology*, 21(7):e3002201, July 2023a.

Gaia Molinaro and Anne G. E. Collins. Human hacks and bugs in the recruitment of reward systems for goal achievement. *Proc. Annu. Meet. Cogn. Sci. Soc.*, 45(45), 2023b.

Gaia Molinaro and Anne G. E. Collins. A goal-centric outlook on learning. *Trends Cogn. Sci.*, 27 (12):1150–1164, December 2023c.

Stephen Monsell and Guy A Mizon. Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? *Journal of Experimental Psychology: Human perception and performance*, 32(3):493, 2006.

Randall C. O'Reilly. Unraveling the mysteries of motivation. *Trends Cogn. Sci.*, 24(6):425–434, June 2020.

Ben Prystawski, Florian Mohnert, Mateo Tošić, and Falk Lieder. Resource-rational models of human goal pursuit. *Topics in Cogn. Sci.*, 14(3):528–549, 2022.

Robert D Rogers and Stephen Monsell. Costs of a predictible switch between simple cognitive tasks. *Journal of experimental psychology: General*, 124(2):207, 1995.

Darryl W Schneider and Gordon D Logan. Defining task-set reconfiguration: The case of reference point switching. *Psychonomic Bulletin & Review*, 14:118–125, 2007.

Pradyumna Sepulveda, Marius Usher, Ned Davies, Amy A Benson, Pietro Ortoleva, and Benedetto De Martino. Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, 9:e60705, 2020.

Brian M. Sweis, Samantha V. Abram, Brandy J. Schmidt, Kelsey D. Seeland, Angus W. MacDonald, Mark J. Thomas, and A. David Redish. Sensitivity to "sunk costs" in mice, rats, and humans. *Science*, 361(6398):178–181, July 2018.