Investigating the Role of Task Representation Switch Costs in Goal Persistence

Gaia Molinaro*
Department of Psychology & Helen Wills Neuroscience Institute
University of California, Berkeley

Aly Lidayan*
Department of EECS
University of California, Berkeley

Anne G. E. Collins
Department of Psychology & Helen Wills Neuroscience Institute
University of California, Berkeley
*equal contribution

Abstract

Goal pursuit profoundly shapes human cognition. This is often beneficial – for example, by focusing information processing. However, goal-dependent computations occasionally lead to seemingly maladaptive behavior, such as a bias toward goal persistence – the tendency to continue pursuing the current goal even when suboptimal. While various psychological explanations have been proposed for goal persistence, the underlying cognitive mechanisms remain unclear. Here, we explore one potential hypothesis. We posit that switching goals incurs the computational cost of reconfiguring internal representations, e.g., to new stimulus-action mappings, and that this cost motivates persisting in a goal even when another is more valuable. Thus, we predict that 1) higher reconfiguration costs will increase persistence bias, and 2) given the choice, participants will prefer switching to lower-cost goals. To test these predictions, we developed a task where participants chose between competing goals, and sudden changes in the task dynamics encouraged goal abandonment toward more valuable options. Crucially, some goals shared the same action selection rule, while others required different rules, allowing us to test how similar vs. different internal representations influenced goal persistence. Across two experiments (N = 125), we replicated goal persistence biases and found that stimulus-action representation switch costs influence both goal persistence and selection preferences. However, they only accounted for a small part of the total goal persistence, suggesting that such biases are predominantly driven by other factors.

Keywords: persistence, goals, behavioral bias, cognitive modeling, switch cost, representation

Introduction

From the dedicated athlete who never misses a training session to the serial entrepreneur who pursues one venture capitalist after another, our society often praises goal persistence. However, people frequently continue pursuing goals long after it has become suboptimal. Failing to disengage from unfavorable objectives that would better be abandoned is a well-established phenomenon known as the "sunk cost fallacy" (Holton et al., 2024; Sweis et al., 2018). This tendency also manifests itself in simpler settings where, if free to choose, people tend to avoid switching tasks (Arrington & Logan, 2005). Understanding why maladaptive goal persistence occurs is crucial for both basic research and practical applications. Clarifying the mechanisms behind goal persistence (or lack thereof) could help treat cases of goal dysregulation occurring in psychopathology or addictive behavior (Hogarth, 2020; Johnson et al., 2012). Moreover, accurate models of human decisions 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).

Several accounts of goal persistence have been offered to date. Under one explanation, sunk cost biases arise to avoid appearing wasteful to others (Arkes & Blumer, 1985). However, evidence for the phenomenon has also been found under non-social conditions Cheng et al., 2025. Alternatively (though not mutually exclusively), "stubborn goals" may have evolved thanks to their generally adaptive roles, helping allocate cognitive resources efficiently, shielding systems from interference, and consistently driving motivation (Holton et al., 2024, 2025; Mittelstädt et al., 2018). While setting a goal often causes a beneficial shift in mindset – inducing attentional, representational, and motivational changes that favor its completion (Castegnetti et al., 2021; Frömer et al., 2019; Gollwitzer, 1990; Holton et al., 2024; McDougle et al., 2022; Molinaro & Collins, 2023a, 2023b; O'Reilly, 2020; Sepulveda et al., 2020) – reconfiguring such a state is computationally and behaviorally expensive (Grahek et al., 2023; Gu et al., 2017; Monsell & Mizon, 2006). Therefore, people's bias for goal persistence could partly stem from representation switch costs, causing people to persist more when switch costs are higher and, when they do switch, to prefer goals with lower costs. For example, imagine a pastry chef who realized their muffins are not turning out the way they had envisioned. In the first place, they might be resistant to giving up on their progress. After deciding they are better off starting from scratch with a new recipe, they might be more likely to opt for cupcakes than croissants, the former requiring a lesser switch in mental frame and activities. They may also persist for longer if cupcakes are not a viable alternative, compared to when they are. Here, we test both of these predictions. Note that in the present context, a task representation refers to a specific configuration of the cognitive system required to perform a task (Schneider & Logan, 2007), namely stimulus-action associations (also called a "task-set" Monsell and Mizon, 2006).

To test our hypotheses, we developed an experimental paradigm inspired by Holton et al. (2024) in which participants chose among competing goals on each trial. We designed the task so that in some trials, persisting with the current goal would yield a more delayed reward compared to

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

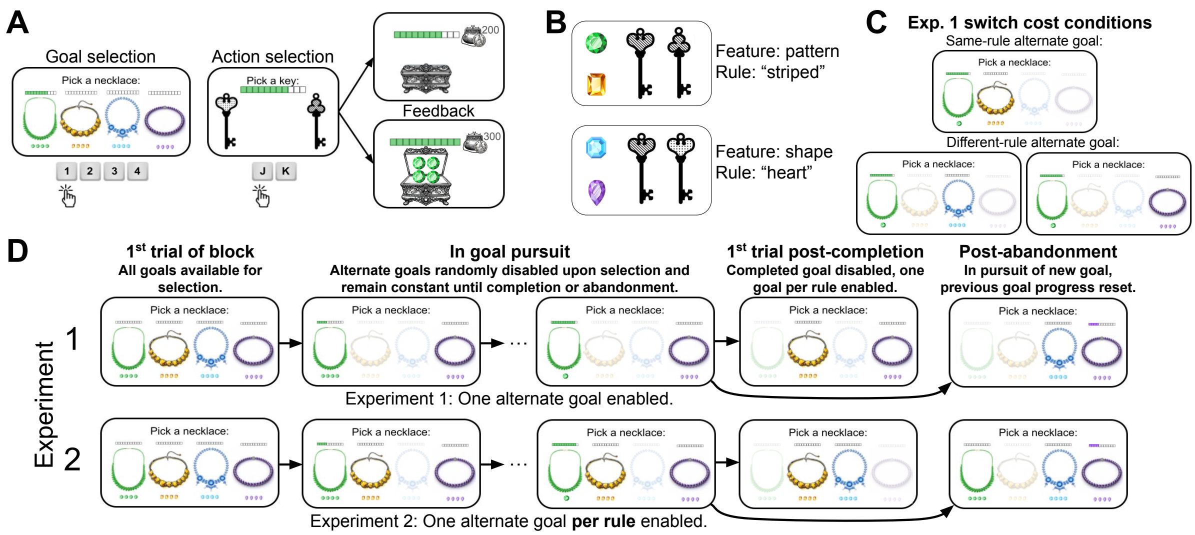

Figure 1: **Experimental design.** (A) Trial structure. On each trial, participants viewed their progress and the number of gems offered for each necklace, and chose a necklace (goal). Next, they chose a key (action), characterized by a shape and pattern. Finally, participants were shown a treasure chest – open and containing gems if they chose the correct key, closed otherwise – the updated progress on the chosen necklace, and the total rewards earned thus far. (B) Example rules. For each participant, one rule for each feature (shape and pattern) was randomly sampled and assigned to gem colors such that goals shared rules in pairs. (C) Example Exp. 1 switch cost conditions under the rules in panel B. In the same-rule condition, the only other available goal shared a rule with the current goal, so goal abandonment did not involve rule switching. In the different-rule condition, switching to the alternative goal implied switching rules. (D) Goal mechanics. All goals were initially available. Upon selecting a new goal, a random set of alternate goals was disabled. This set remained constant during goal pursuit until the current goal was completed or abandoned. In the first trial post goal completion, the completed goal was disabled, and one goal for each rule was enabled. Abandoning the current goal reset any ongoing progress and entered pursuit of the new goal. In Exp. 1, only one alternate goal was enabled during pursuit, creating the two conditions shown in panel C. In Exp. 2, one different-rule goal was disabled during goal pursuit to equate the number of alternate goal options per rule.

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 measured task representation switch costs as individuals' increase in reaction time (RT) following a switch in rules, relative to switching goals without switching rules. If costs associated with task representation switches underlie people's tendency to persist, participants should be more likely to abandon suboptimal goals when the alternative does not require switching rules, compared to when it does. Moreover, if people consider these costs when selecting goals, they should prefer switching to a goal that shares a rule with the previous goal, compared to one that does not. In two separate studies, we find strong goal persistence behavior, as well as evidence in favor of our hypotheses. However, switch costs only accounted for a small fraction of goal persistence. We discuss alternative mechanisms for goal persistence.

Results

Experimental setup

Based on an existing paradigms (Holton et al., 2024), we developed a novel experimental protocol, detailed in the Methods, in which participants were rewarded for completing necklaces (Figure 1). In brief, on each trial, participants first selected one of four possible necklaces (goals) as their target (Figure 1A, Goal selection). Participants were then presented with two keys, each characterized by a shape (heart or clover) and a pattern (striped or dotted). Necklaces were organized into two pairs sharing a stimulus-action rule (Figure 1B), with one pair's rule based on the pattern and the other based on the shape. Selecting the correct key yielded a specified "offer quantity" of gems matching the chosen necklace's color, incrementing progress toward completing the goal. Once all required gems for the active necklace had been collected, participants received a reward of 100 coins added to their "purse" (Figure 1A, Feedback), and their progress was reset at the next trial. To study rule switching preferences after goal completion, we disabled both the goal that was just completed and one additional goal, such that there was one new goal of each rule available (Figure 1D, 1st trial post-completion). To incentivize participants to occasionally abandon goals during pursuit, the offer quantity for the in-progress necklace could suddenly decrease to 1 gem ("depletion"; Figure 1D, In goal pursuit), remaining at that level until necklace completion, while the offer quantity for other goals remained unchanged

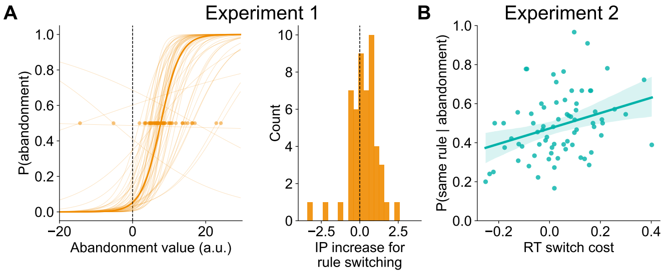

Figure 2: **Behavioral and modeling results.** (A) Left: Probability of goal abandonment as a function of its value, assuming no switch costs. The bold line shows fit across all participants; transparent lines and dots show individual participants and their indifference points (IP) to abandonment. Most participants showed a bias towards persisting, as indicated by IPs above zero. Right: distribution of individuals' change in IPs between same- and different-rule switch cost conditions. For most participants, persistence bias was sensitive to switch cost. (B) Switch costs (differences in mean log RTs for stimulus choices on the first trial after switching to a same- vs. different-rule goal immediately after completion) correlated with the probability of switching to a same-rule goal after abandonment in Exp. 2. Dots indicate individual participants, shading indicates the correlation's 95% CI.

– often making abandonment the optimal decision. For instance, if the offer depletes from 5 gems to 1 gem while the current progress is only 5/20 gems, persisting with the current goal would take at least 15 trials to receive the next reward, whereas abandoning could let the participant complete a different goal in as few as 4 trials. Abandoning a goal restored the offer quantity but reset any progress (Figure 1D, Post-abandonment). We disabled subsets of the alternate goals during goal pursuit (Figure 1D, In goal pursuit). In Exp. 1, to test if persistence biases were sensitive to switch cost, exactly one alternate necklace was enabled, creating two switch-cost conditions depending on whether the alternate necklace shared a rule with the current goal (Figure 1C). In Exp. 2, to measure goal selection preferences at abandonment, one alternate necklace per rule was enabled during pursuit.

Action selection reveals goal switching costs

After applying our exclusion criteria (see the Methods), the average rate of selecting the correct action was extremely high (Exp. 1 M = 0.94 ± 0.006; Exp. 2 M = 0.94 ± 0.005). Participants abandoned their goals several times throughout the task (Exp. 1 M = 21.73 ± 1.36; Exp. 2 M = 22.97 ± 1.09). In the majority of cases, abandonment choices were made following gem depletion (Exp. 1 M = 93% ± 0.02, Exp. 2 M = 91 ± 0.02). Compared to trials in which goals remained consistent, switching goals resulted in significantly slower action selection RTs, whether the switch was initiated voluntarily (Exp. 1 $t(51) = -8.25$, $p < 0.001$; Exp 2. $t(72) = -11.75$, $p < 0.001$) or forced by task design after the previous goal was completed (Exp. $t(51) = -6.72$, $p < 0.001$; Exp. 2 $t(72) = -8.64$, $p < 0.001$). Having found evidence that participants adequately understood the task and showed measurable goal switch costs, we set out to test individuals' goal persistence biases and their association with task representation switch costs.

Participants display strong goal persistence biases

Participants displayed strong persistence biases with respect to goals in our task. To show this, we followed Holton et al. (2024) and modeled the value of goal abandonment in terms of the expected reduction in steps needed to complete a goal upon abandoning compared to persisting with the current one (see the Methods section for details). We fit this model via a mixed-effects logistic regression predicting abandonment choices from abandonment values with random intercepts and slopes for each participant. The indifference point (IP) corresponds to the abandonment value at which abandoning or persisting is equally likely. If participants behaved optimally, their IPs would be at zero, meaning they would only be more likely to persist with a current goal with a lower per-trial offer quantity if they expected to earn a reward in fewer trials than if they switched to an alternative goal (higher-paying, but requiring starting from scratch). Instead, IPs higher than zero would translate into a negative coefficient for the intercept in the mixed-effects model (Methods) – as we found in both Exp. 1 ($\beta = -2.89 \pm 0.26$, $p < 0.001$) and Exp. 2 ($\beta = -2.49 \pm 0.26$, $p < 0.001$). In other words, participants persisted with their current goal beyond what was optimal given the task constraints.

Task representation switching has a small effect on persistence bias

We next investigated whether task representation switch costs played a role in goal persistence bias by analyzing the data from Exp. 1, where we manipulated whether goal abandonment involved rule switching. First, we verified that rule switching increased the representation switch costs of goal abandonment, as measured by a participant's (log) RTs for action selection in the first trial after abandoning a goal (Bustos et al., 2024). We found that across participants, the RTs were significantly greater when abandoning to a goal with a different rule (M = 6.53 ± 0.03) than one with the same rule as the previous goal (M = 6.57 ± 0.02, $t(51) = -2.34$, $p = 0.023$). Nonetheless, participants showed significant persistence biases even when the alternative goal did not require any representation switch ($\beta = -3.03 \pm 0.31$, $p < 0.001$). To test whether persistence biases were sensitive to individual differences in representation switch costs, we added participants' RT costs as a fixed effect in our logistic mixed-effects model (Methods), and found it to be significant ($\beta = -0.74 \pm 0.34$, $p = 0.029$), where the negative coefficient indicates that the persistence bias increased with the cost. This model fit goal persistence choices better than one without RT costs as a factor (likelihood-ratio test $\chi^2(1) = 4.62$, $p = 0.032$) and had a slightly lower AIC (2574.8 vs. 2577.4). However, the effect was small relative to the overall persistence bias: the mean indifference point (IP) across participants was at an abandonment value of 7.4, while the mean magnitude of the shift in IP between switch cost conditions was only 1.3 (Figure 2A), and the option to switch to a goal with the same rule (compared to a different one) only decreased the IPs by an average of 2.2% of the subjective cost.

Task representation switch costs correlate with goal switching preferences

To test whether task representation switch costs impact which goal participants chose to abandon current goals *for*, we analyzed data from Exp. 2, where participants could freely choose between one alternate goal of each rule during goal pursuit.

First, we again verified that rule switching incurred a cost by comparing average log RTs in the first trial after switching to a different- vs. a same-rule goal. The measurable costs were less than in Exp. 1, as participants who would have incurred the largest costs may have always chosen to abandon to same-rule goals, but we still found a small significant cost of task representation switching (same-rule M = 6.55 ± 0.02, different-rule M = 6.58 ± 0.02, $t(72) = -2.41$, $p = 0.018$).

Contrary to our prediction, we found no overall preference across participants for selecting goals with the same rule (M = 0.49 ± 0.02, $t(72) = -0.62$, $p = 0.537$). However, we hypothesized that individual differences in participants' ability to efficiently transition between different goal-dependent representations could impact their goal selection. Supporting this prediction, participants' preferences for same-rule goal selection were sensitive to their representation switching RT costs ($R(72) = 0.32$, $p = 0.006$; Figure 2B).

Taken together, these results indicate that rule switching in our experiments incurred a small but significant RT cost, which led to individual differences in both *when* to abandon a goal and *which* alternative to abandon *for*. However, the task representation switch costs only explained a small fraction of the large overall goal persistence bias observed.

Discussion

Whether in admiration for an artist spending years mastering their piece or respect for a scientist engaging in decade-long research, goal persistence is often highly praised. However, the tendency to persevere can lead people to pursue goals that would better be abandoned (Arkes & Blumer, 1985; Hammond et al., 1998). Understanding the root causes of goal persistence is key to steering goal pursuit towards adaptive outcomes and away from maladaptive habits. Here, we set out to investigate the role of task representations in goal persistence, proposing that part of the resistance to abandoning current pursuits could be attributed to the cognitive costs of remapping goal-contingent stimulus-action associations.

In a novel experimental setting, we replicate the finding that people persist in their goal pursuits long after it becomes suboptimal to do so, compared to switching to alternative goals. Similar biases have previously been observed in voluntary task switching tasks, where participants freely select which of multiple tasks to perform and, despite being instructed to choose randomly, show a consistent bias to repeat the same task on consecutive trials (Arrington & Logan, 2004, 2005). These experimental designs, however, usually offer no advantage to switching and no notion of progress with the current goal relative to others. Addressing both points, our results more closely match evidence for the "sunk cost fallacy" found in settings where abandoning the current goal is inherently suboptimal for task completion (Holton et al., 2024; Sweis et al., 2018). Nevertheless, our study is uniquely positioned to relate goal persistence biases with task representation switch costs. Consistent with the finding that task switch costs increase as task rules and objectives become more dissimilar (Bustos et al., 2024; Grahek et al., 2023), our data shows slower reaction times after switching to a task requiring the same stimulus-action mapping compared to a different one. We also find some evidence that differences in task representation switch costs correlate with goal persistence bias, suggesting that individuals' decisions to persist or abandon may be, at least in part, resource-rational (Prystawski et al., 2022). However, we did not observe that such switch costs could fully explain the large goal persistence biases we measured. This result aligns with the finding that reducing switch costs does not always lead to a corresponding reduction in task repetition biases (Mittelstädt et al., 2018), and suggests that factors other than task representation consistency more strongly contribute to goal persistence biases.

One possible explanation is that representation switches do play a role in persistence biases, but operate at the level of the goal as a whole, rather than each of its components. Some evidence suggests that goal and rule selection proceed in sequence, rather than in parallel – that is, control systems first choose a goal and then activate the rule associated with it Rubinstein et al., 2001. This view fits the perspective that goal-directed behavior mainly gravitates around two different mindsets, goal-selection vs. goal-engaged – the first associated with initial openness to all possible goals and the latter focusing resources on a single objective after one is selected (Gollwitzer, 1990; O'Reilly, 2020; O'Reilly et al., 2014). In line with this framework, previous experiments found strong preferences for goal consistency even after goal completion, when sunk costs are no longer relevant (Arrington & Logan, 2005; Molinaro et al., n.d.). An alternative perspective is that people evolved goal persistence biases to avoid re-evaluation, which is adaptive when past experiences are highly predictive of the future and only becomes suboptimal in cases (such as our experimental paradigm) where sudden changes in the task dynamics occur (Aenugu & O'Doherty, 2025; Krämer, 2017). People may also over-persist to preserve self-consistency and avoid cognitive dissonance or feelings of incompetence (Dörner, 2003; Festinger, 1959). Any of these mechanisms suggests high costs for goal abandonment, but potentially lower costs for the subsequent goal selection, which could explain why, in our experimental setting, participants were reluctant to abandon a goal but showed no strong preferences for alternative goals with a shared rule. Indeed, goal persistence may be primarily driven by switch costs we did not directly measure, such as recalling goal-rule associations – which would be equivalent across all alternative goals. Finally, social accounts have proposed that goal persistence stems, at least in part, from a desire for consistency, which is beneficial in social settings where coordinating with others is facilitated by inferring their goals (Cheng et al., 2023), but keeps manifesting even when individuals act on their own and unobserved (Cheng et al., 2025).

We also recognize some limitations of our experimental approach. For example, the same goals were always associated with the same keyboard key to avoid participant confusion. As a result, goal persistence could have been partly driven by motor decisions to avoid the effort of pressing a different button or habitual processes. However, experimental designs either measuring or eliminating this confound still showed evidence for goal persistence beyond such simple explanations (Bustos et al., 2024; Holton et al., 2024). Moreover, it is possible that the representation switches we implemented were not complex enough to induce a strong effect. In developing our experimental paradigm, we faced a trade-off between making the rules themselves easy to learn (to ensure they knew the correct goal-rule mapping) and the action representations difficult to alternate (to cause a switching cost). Future studies may improve the current design, e.g., by keeping rules simple but introducing perceptually more challenging tasks to increase the switch costs specific to action representations. Such a manipulation could induce larger effects of representation switch costs on goal persistence; however, even in our current design, persistence biases were comparatively large, suggesting other factors play a role in people's susceptibility to sunk costs. Follow-up work could also further investigate the role of individual differences in goal persistence's sensitivity to representation switch costs, e.g., by collecting additional data about participants' cognitive flexibility or socio-emotional dimensions. Our model could also be extended, e.g., by including proportional goal progress or number of trials invested in the current goal as additional predictors.

In sum, we replicate the finding that people have strong biases to keep pursuing current goals even when suboptimal. Contrary to our expectations, the costs associated with task representation switches only played a small role in goal persistence. These results suggest that proposing alternative goals may be an ineffective strategy to correct maladaptive goal choices, even if switching would require minimal changes in behavior. Instead, operating directly on the current goal itself, e.g., by interrupting existing cycles of goal pursuit, might be a more promising way to help people abandon detrimental goals for more empowering ones.

Methods

Experimental procedure

In this task (Figure 1), participants collected gems to complete different colored necklaces (goals). On each trial, participants were first given up to 8 s to select one of four possible necklaces. For each, the size (number of gems needed to complete it), goal progress (number of gems already collected), and offer quantity (number of gems obtainable at that trial) were displayed. Participants were then presented with two keys, each characterized by a shape (heart or clover) and a pattern (dotted or striped). Only one key feature was relevant for obtaining gems of a particular color. The four necklaces were created in two pairs with shared action selection rules: one pair required selecting keys based on pattern, and the other based on shape (for example, green and yellow required selecting the stripe-patterned key; blue and purple the heart-shaped key, as shown in Figure 1B). Necklace positions and rules were randomized across participants. Selecting the correct key yielded gems matching the chosen necklace's color, incrementing goal progress by the offer quantity, while selecting the incorrect key or failing to choose within 1.2 s yielded zero gems and left goal progress unchanged (Figure 1A). Upon collecting a number of gems equal to or higher than the necklace size, participants received 100 points, and progress was reset to 0 (Figure 1D, post-completion).

To induce participants to occasionally switch away from initiated goals, the offer quantity for the current in-progress necklace could suddenly decrease to one gem per trial, remaining at this level until necklace completion or switching to a new necklace, but switching reset progress on the previous necklace (as shown in Figure 1D, Post-abandonment). Baseline offer quantities and necklace sizes were equal across gem colors and stable within each of the three blocks (80 trials each), with block order pseudo-randomized across participants. Baseline offers were 5, 6, or 7 gems, with corresponding sizes 20, 24, and 21, respectively. Depletion occurred during roughly 30% of goal pursuits, occurring at goal progress thresholds ranging from 25%-66%, leading to abandonment values ranging from -17 to 13.4.

To test our hypotheses about task representation switch costs, subsets of the alternate necklaces were disabled during goal pursuit (Figure 1D, In goal pursuit). To test if persistence biases were sensitive to switch cost, in Exp. 1, exactly one alternate necklace was enabled during pursuit, with the same rule as the current necklace in 50% of the trials and a different rule in the other 50%, corresponding to two different switch costs (Figure 1C). To measure goal selection preferences at abandonment, in Exp. 2, one alternate necklace per rule was available during pursuit. In addition, in both experiments, the completed necklace and one additional necklace were disabled in the first trial post-completion, such that one new necklace for each rule was available. This allowed us to study the effect of task switching on goal selection after goal completion.

Participants underwent extensive training prior to the main task. They were first taught the interface and dynamics of the task, including the fact that gem depletions occurred randomly and that the rate of depletion did not depend on the color. Participants then 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, to avoid potential order effects in learning the rule for each necklace. Participants repeated each practice block until they reached 80% accuracy, for a maximum of 3 times. This ensured that, by the end of practice, each participant had learned the rules associated with each necklace.

Participants

Participants residing in the USA aged 18–40 were recruited from Prolific (https://www.prolific.com/) and compensated with $12 for approximately one hour of their time. No additional performance bonuses were disbursed. Following Hughes et al. (2014), we excluded participants for achieving less than 80% accuracy in the key selection task. Following Xia et al. (2021), elbow plots were used to determine cutoffs for participant exclusion based on the number of trials where participants failed to select a necklace within the allotted time (greater than 5) or abandoned an in-progress goal (greater than 65), all indicating poor task comprehension or compliance. In Exp. 1, we collected data from a total of 132 participants. We excluded 23 for achieving less than 80% accuracy in the key selection task, 4 for failing to select a necklace within the allotted time on more than 5 trials, and 1 for abandoning their goal more than 65 times. In Exp. 2, we collected data from 126 participants and excluded 19 for achieving less than 80% accuracy in key selection, 3 for failing to select a necklace on more than 5 trials, and 1 for abandoning more than 65 times. In addition, we could only calculate task switching RT costs for participants who abandoned their goal for same- and different-rule necklaces at least once each, ruling out 27 and 24 of the remaining participants for Exp. 1 and 2, respectively. Finally, we restricted our analyses to participants who abandoned and persisted at least twice each on depleted trials for each switch cost condition, in order to reliably estimate persistence bias on depleted goals and how it is affected by switch cost. Exp. 1 had two switch cost conditions, leaving 52 subjects with sufficient data, while Exp. 2 had one switch cost condition, leaving 73. Although our other analyses, e.g., of goal selection preferences, do not require this final restriction, we note that our findings remain unchanged if we relax it, and we report results on the same set of 125 subjects across analyses for consistency. The University of California, Berkeley IRB approved the experimental procedure (protocol #2024-04-17357).

Modeling

Following Holton et al. (2024), we modeled the value of abandonment as the estimated decrease in the number of steps until goal completion if an alternative necklace is picked over the current one, $\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}$. The subjective value of abandonment SV_{abandon} and probability of abandonment P_{abandon} were modeled as:

$$SV_{\text{abandon}} = \beta_0 + \beta_1(\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}). \tag{1}$$

$$P_{\text{abandon}} = \frac{1}{1 + e^{-SV_{\text{abandon}}}}. \tag{2}$$

We only analyzed choices made in trials where the goal offer quantity was depleted, since alternative goals offer no advantage compared to the current goal before depletion [1]. To estimate $\hat{T}_{\text{current}}$, we used the fact that if it was optimal to persist with a depleted goal, this would remain the case until goal completion, since the depleted offer quantity could not further decrease and the goal progress could 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 goal completion assuming no switches or incorrect actions). Since all necklaces shared a constant baseline offer quantity and size throughout the block, we estimated $\hat{T}_{\text{alt}}$ as the expected number of trials participants needed to complete a goal from scratch in the current block assuming perfect choice accuracy, which we calculated by dividing the number of trials with correct key choices by the number of necklaces they completed in that block.

To investigate whether individual differences in task representation switch costs impacted persistence biases in Exp. 1, we added each participant's (p) switch cost C^p_{switch} to the estimated subjective value of abandonment:

$$SV'_{\text{abandon}} = \beta_0 + \beta_1(\hat{T}_{\text{current}} - \hat{T}_{\text{alt}}) + \beta_2 C^p_{\text{switch}}. \tag{3}$$

On trials where the current and alternate goals did not share rules, abandoning incurred an additional rule-switching cost of $C^p_{\text{switch}} = RT^p_{\text{diff}} - RT^p_{\text{same}}$ (the individual change in mean log RT when abandoning for different- vs same-rule goals). On trials where the alternate and current goals shared the same rule, the cost was opposite $C^p_{\text{switch}} = RT^p_{\text{same}} - RT^p_{\text{diff}}$ (reflecting facilitation compared to a different-rule goal).

We fit a mixed-effects logistic regression model predicting abandonment choices with random intercepts and abandonment value slopes for each participant, while the effect of C_{switch} was modeled with a fixed β_2 coefficient. We combined the fixed and random effects to calculate the indifference point (where P_{abandon} = 0.5) of each participant p:

$$\hat{T}_{\text{current}} - \hat{T}_{\text{alt}} = -(\beta^F_0 + \beta^R_0 + \beta^F_2 C^p_{\text{switch}})/(\beta^F_1 + \beta^R_1), \tag{4}$$

where superscripts F and R denote fixed and random effects, respectively. The indifference points (IPs) for SV_{abandon} follow the same equation without the C^p_{switch} term.

Acknowledgments

This project was funded by a Starting Grant to G.M. and A.L. from the University of California, Berkeley Institute of Cognitive and Brain Sciences. A.G.E.C. and G.M. are also partly

[1]Participants sometimes abandoned in the *second* trial after offer depletion, a suboptimal choice likely reflecting hurried key presses before registering the change. Thus, if the goal was abandoned in the second trial, we excluded the first trial from our modeling analysis.

supported by NSF Grant 2336466 awarded to A.G.E.C. A.L. is also partly supported by the Center for Human-Compatible AI and the Fannie and John Hertz Foundation. We thank Stuart J. Russell for useful comments on an earlier draft of this work.

References

Aenugu, S., & O'Doherty, J. P. (2025). Building momentum: A computational account of persistence toward long-term goals. *PLOS Computational Biology*, *21*(5), e1013054.

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

Arrington, C. M., & Logan, G. D. (2004). The cost of a voluntary task switch. *Psychological science*, *15*(9), 610–615.

Arrington, C. M., & Logan, G. D. (2005). Voluntary task switching: Chasing the elusive homunculus. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 683.

Bobu, A., Scobee, D. R., Fisac, J. F., Sastry, S. S., & Dragan, A. D. (2020). Less is more: Rethinking probabilistic models of human behavior. *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, 429–437.

Bustos, B., Mordkoff, J. T., Hazeltine, E., & Jiang, J. (2024). Task switch costs scale with dissimilarity between task rules. *Journal of Experimental Psychology: General*, *153*(7), 1873.

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

Cheng, S., Zhao, M., Tang, N., Zhao, Y., Zhou, J., Shen, M., & Gao, T. (2023). Intention beyond desire: Spontaneous intentional commitment regulates conflicting desires. *Cognition*, *238*, 105513.

Cheng, S., Zhu, J., Zhou, J., Shen, M., & Gao, T. (2025). Spontaneous coordination with self-commitment: How the presence of others alters the strength, goal and timing of commitment.

Dörner, D. (2003). The mathematics of emotions. *The logic of cognitive systems: proceedings of the Fifth International Conference on Cognitive Modeling; ICCM-5 2003*.

Festinger, L. (1959). Cognitive dissonance. *New York*.

Frömer, R., Dean Wolf, C. K., & Shenhav, A. (2019). Goal congruency dominates reward value in accounting for behavioral and neural correlates of value-based decision-making. *Nat. Commun.*, *10*(1), 4926.

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

Grahek, I., Leng, X., Musslick, S., & Shenhav, A. (2023). Control adjustment costs limit goal flexibility: Empirical evidence and a theoretical account. *bioRxiv*, 2023–08.

Gu, S., Betzel, R. F., Mattar, M. G., Cieslak, M., Delio, P. R., Grafton, S. T., Pasqualetti, F., & Bassett, D. S. (2017). Optimal trajectories of brain state transitions. *NeuroImage*, *148*, 305–317.

Hammond, J. S., Keeney, R. L., & Raiffa, H. (1998). The hidden traps in decision making. *Harvard business review*, *76*(5), 47–58.

Hogarth, L. (2020). Addiction is driven by excessive goal-directed drug choice under negative affect: Translational critique of habit and compulsion theory. *Neuropsychopharmacology*, *45*(5), 720–735.

Holton, E., Grohn, J., Ward, H., Manohar, S. G., O'Reilly, J. X., & Kolling, N. (2024). Goal commitment is supported by vmPFC through selective attention. *Nature Human Behaviour*, 1–15.

Holton, E., Niv, Y., & O'Reilly, J. X. (2025). The adaptive value of stubborn goals. *Trends in Cognitive Sciences*.

Howes, A., Jokinen, J. P., & Oulasvirta, A. (2023). Towards machines that understand people. *AI Magazine*, *44*(3), 312–327.

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior research methods*, *46*(3), 702–721.

Jacob, A., Gupta, A., & Andreas, J. (2024). Modeling boundedly rational agents with latent inference budgets. *International Conference on Learning Representations*.

Johnson, S. L., Fulford, D., & Carver, C. S. (2012). The double-edged sword of goal engagement: Consequences of goal pursuit in bipolar disorder. *Clinical Psychology & Psychotherapy*, *19*(4), 352–362.

Krämer, A. (2017). Demystifying the" sunk cost fallacy": When considering fixed cost in decision-making is reasonable. *Journal of Research in Marketing (ISSN: 2292-9355)*, *7*(1), 510–517.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., & Griffiths, T. L. (2025). Large language models assume people are more rational than we really are. *ICLR*.

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

Mittelstädt, V., Dignath, D., Schmidt-Ott, M., & Kiesel, A. (2018). Exploring the repetition bias in voluntary task switching. *Psychological Research*, *82*(1), 78–91.

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

Molinaro, G., & Collins, A. G. E. (2023b). A goal-centric outlook on learning. *Trends Cogn. Sci.*, *27*(12), 1150–1164.

Molinaro, G., Lidayan, A., & Collins, A. (n.d.). Investigating the role of representation switching costs in goal persistence bias. *Second Workshop on Representational Alignment at ICLR 2025*.

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

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

O'Reilly, R. C., Hazy, T. E., Mollick, J., Mackie, P., & Herd, S. (2014, April). Goal-driven cognition in the brain: A computational framework.

Prystawski, B., Mohnert, F., Tošić, M., & Lieder, F. (2022). Resource-rational models of human goal pursuit. *Topics in Cogn. Sci.*, *14*(3), 528–549.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of experimental psychology: human perception and performance*, *27*(4), 763.

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

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., & De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, *9*, e60705.

Sweis, B. M., Abram, S. V., Schmidt, B. J., Seeland, K. D., MacDonald, A. W., Thomas, M. J., & Redish, A. D. (2018). Sensitivity to "sunk costs" in mice, rats, and humans. *Science*, *361*(6398), 178–181.

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.