Trends in Cognitive Sciences

CellPress OPEN ACCESS

Review

The adaptive value of stubborn goals

Eleanor Holton [1,2,*], Yael Niv [1,3], and Jill X. O'Reilly [2,*]

Humans exhibit a striking tendency to persist with chosen goals. This strong attachment to goals can often appear irrational – a perspective captured by terms such as perseverance or sunk-cost biases. In this review, we explore how goal commitment could stem from several adaptive mechanisms, including those that optimise cognitive resources, shield decisions from interference, and scaffold motivation in the absence of accessible reward signals. We propose that these computational considerations have important implications for algorithmic architectures supporting decision making, including separate algorithms for goal selection and implementation, and for monitoring ongoing goals versus alternative sources of reward. Finally, we discuss how a variety of mechanisms supporting goal commitment and abandonment could relate to dimensions affected in mental health.

Highlights

Humans show strong attachment to goals they have selected.

While often framed as a maladaptive bias, we outline three adaptive functions leading to stable goals: efficient cognitive resource allocation, shielding from interference, and scaffolding motivation in the absence of immediate and tangible reward signals.

These considerations shape the algorithmic architectures that support naturalistic goal pursuit, such as the mechanisms that select, implement, and revise goals.

Understanding how these different mechanisms become miscalibrated may have important implications for characterising alterations in goal commitment affecting mental health.

Commitment to goals

Humans are notoriously reluctant to abandon goals [1]. We finish races despite bad injuries, persist in the same queue when another lane is moving faster, and get trapped in hours of gaming to progress just one more level. This seemingly inflexible commitment to the goals we have chosen is often considered irrational, and indeed in laboratory tasks, over-commitment or perseveration usually translates into worse performance. To account for such irrational attachment to goals, computational models of decision making often include parameters, labelled 'perseverance bias', 'over-staying', or 'stickiness', treated as a nuisance measure or bias [2–6]. Yet across the lifetime, costs and benefits are not assessed in terms of immediate outcomes – whether points in a laboratory task or minutes lost in a queue – but in relation to broader constraints. Agents must balance limited resources, make choices with long-term consequences, and act in complex social environments. These pressures can make goal persistence adaptive in real-world settings, even if, when observed in isolation, the same behaviours appear suboptimal.

Here we focus on three pressures that favour stable goals: management of cognitive and energetic resources, shielding agents from interference (e.g., in the form of short-term rewards), and scaffolding motivation in the absence of external reward signals (Figure 1). An understanding of these adaptive reasons for goal commitment (i.e., reasons at the 'computational level', in Marr's terminology [7]) is critical for characterising the assortment of behaviours currently grouped as irrational forms of 'perseveration'.

This adaptive perspective should also inform our models of how people make decisions. For example, algorithmic approaches that separate goal selection from pursuit, or that monitor chosen goals differently from non-goal rewards may better capture naturalistic behaviour. Differences in how these distinct algorithms are tuned may also capture basic dimensions of mental health.

Defining goals and goal stability

Before we dive in, we would like to clarify our use of the term 'goal'. In daily life, we are faced with countless sources of potential reward, most of which take time to attain, and whose value is often uncertain. Should we clean the kitchen, meet friends, or apply for new jobs? Given these diverse

[1]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA
[2]Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK
[3]Department of Psychology, Princeton University, Princeton, NJ 08540, USA

*Correspondence: eleanor.holton@princeton.edu (E. Holton) and jill.oreilly@psy.ox.ac.uk (J.X. O'Reilly).

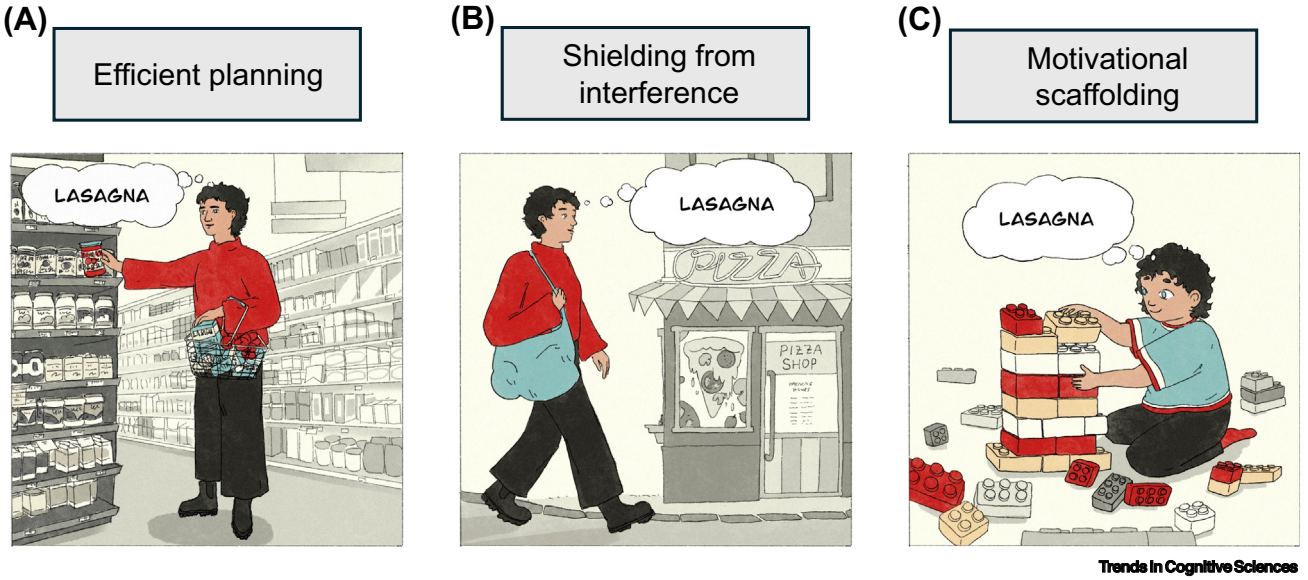

Figure 1. Adaptive mechanisms leading to stable goals. (A) Goal commitment could arise from efficient use of cognitive resources – for example, allowing cognitive resources to be used for implementing the goal (e.g., shopping for lasagna ingredients), rather than continuously deliberating about alternative goals (e.g., when faced with ingredients for other possible recipes). (B) Mechanisms that limit reconsideration of alternative options after goal selection may help insulate decision making from transient fluctuations in preferences, for example, the short-term temptation of getting a take-out. These mechanisms could shield long-term goals from interference while leading to goal commitment. (C) Beyond their role in constraining the space of options, goal selection can generate motivational signals that could shape behaviour in the absence of external rewards. This would lead to stable goals by making pursuit of the goal intrinsically valuable. For example, the motivation to complete self-imposed goals during play – such as a child carefully assembling a Lego lasagne – could promote learning and exploration [12]. In turn, this 'scaffolding' role of goals extends into adulthood, particularly when the connection between behaviour and the target outcome requires support through an intermediate reward signal. Artwork by Magdalena Adomeit.

pulls, the ability to select specific outcomes as current 'goals' and be motivated to pursue them is fundamental to how we structure life [8–12] (Box 1). Here, we define goals as mental representations of desired future outcomes that guide sustained behaviour [13,14]. These are distinct from the specific actions involved in achieving the goal or the goal's utility.

Goals – once selected – exhibit a remarkable stability. By this, we mean that the mere fact that an action aligns with a selected goal increases its likelihood of being chosen. Empirically, this is identified by comparing how choices between the same options differ if probed before versus after goal selection [15], or by modelling the subjective value of offers congruent with selected goals versus alternative sources of value [16–19]. Importantly, this approach has uncovered suboptimal biases toward goal persistence in laboratory paradigms, relative to maximising study reward, or maximising estimated reward from a reinforcement learning perspective [16–20]. The question therefore arises, why do we observe such strong, seemingly maladaptive commitment to goals?

Box 1. Goals and intentions

We define goals as mental representations of desired future outcomes that guide sustained behaviour [13,14]. This concept of goals bears an important resemblance to the concept of 'intentions' as understood by many philosophers [44]. Intentions here are mental states that play a causal role in guiding action. Under this account, intentions are defined as having two key features. First, they are controlling in the sense that, once formed, they tend to guide behaviour automatically unless interrupted [44]. Second, intentions are stable in the sense that they resist re-evaluation. Specifically, a higher amount of contradictory evidence would be required to trigger re-evaluation of an intention than the evidence required to form an intention in the first place [44,49]. Philosophers contend that forming an intention (e.g., to make lasagne) affects actions taken in the future (e.g., searching for recipes, buying ingredients), as well as what information is represented (e.g., ignoring an advert for a new pizza place). Notably, others have used the term 'intention' to refer to the specific implementation plan for a goal [81,82], but here we focus on its use to denote a preselected goal state, held fixed during pursuit.

Behavioural economists and psychologists have long been interested in the 'irrational escalation in commitment' to our past decisions [1,21]. Notably, the simplest account of this behaviour – the 'sunk-cost' account – does not invoke the idea of 'goals' but relies only on the concepts of losses and gains [1]. 'Sunk costs' arise when decisions are excessively influenced by past investments of effort, time, or money, which should be irrelevant to considerations of utility since the investment has already been made and cannot be recovered (i.e., is 'sunk'). However, under prospect theory, lost investments carry greater subjective weight than gains of equivalent objective value [22,23], so people persist with a project to recover losses more than they would from a position of no prior investment.

However, 'sunk costs' fail to account for a more general notion of goal commitment because studies show investments of resources are not necessary for eliciting commitment. Often, merely making the decision to pursue a particular goal leads to inflexibility around its reconsideration as illustrated in Figure 2A,B [15,16]. In one recent study, subjects chose between two goals they

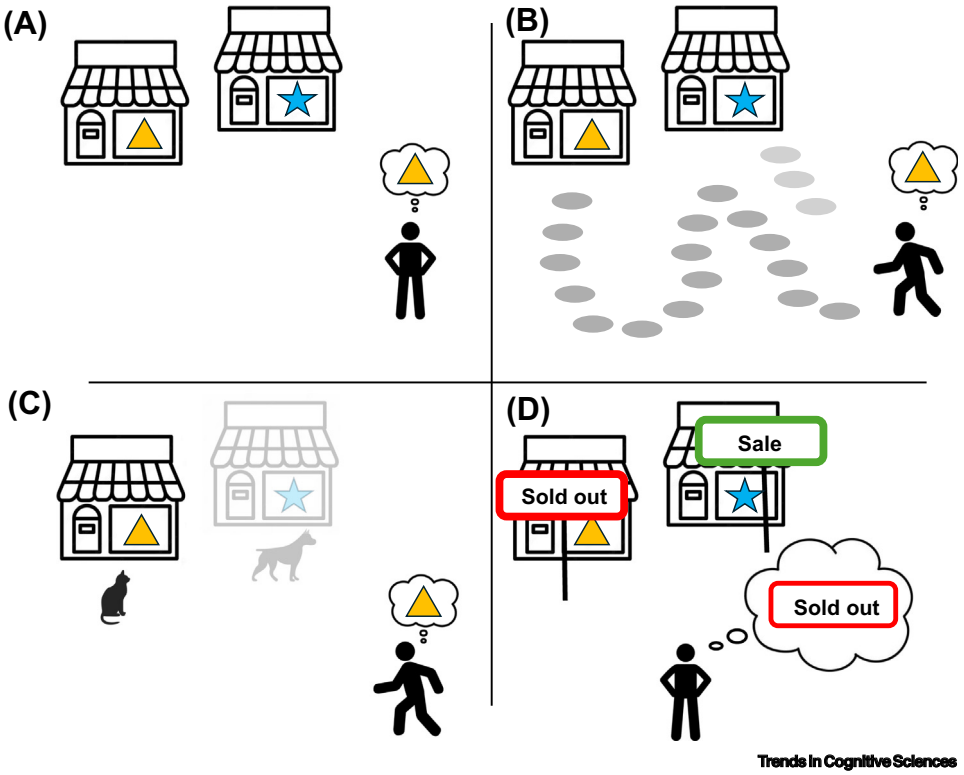

Figure 2. Features of goal commitment. (A) Goal selection and commitment: once a goal is selected, cognition changes in various ways. Here, we illustrate current findings through the example of preselecting a restaurant. (B) Bias toward goal persistence: people are biased toward persisting with selected goals (e.g., Restaurant A) in the face of new information favouring abandonment (e.g., discovering a shorter path to an alternative [16,17], or more effort required to complete the goal [15]). (C) Attentional filtering during goal pursuit: selecting a goal triggers attentional filtering of information irrelevant to the chosen goal (e.g., location of the dog) while enhancing goal-relevant information (e.g., location of the cat). This attentional bias persists outside of decision periods, correlates with individual differences in goal commitment, and builds across goal pursuit [17]. Such filtering may reflect the formation of a simplified 'task construal' that optimises cognitive resources [34,35]. (D) Monitoring the chosen goal: during goal pursuit, people exhibit greater sensitivity to negative changes in their current goal's value ('frustration') than to positive changes in alternative goal values ('temptation'). In a recent study, participants incrementally progressed toward a chosen goal while continuously receiving information about the value of their own and alternative goals (which varied independently) [17]. Although participants should sometimes switch goals, they increasingly ignored information about alternative goals, measured as insensitivity to such information in both decision making and reaction times. However, they remained responsive to devaluations of their current goal, consistent with continued monitoring of the current goal value.

would later have to pursue, with their respective costs presented either before or after selecting the goal [15]. For both adults and children, making an initial selection led to strong perseveration with the same goal after costs were revealed, despite no investment in either goal. In other words, goals become stable even before any investments of time, money, or resources have been made. So where does this commitment come from?

The computational level: three rational reasons for goal stability

What may appear as irrational goal commitment in the context of a single task can be rational when evaluated against broader cost functions. When we consider agents navigating complex decisions with limited resources and long-term repercussions, there are multiple reasons why goal persistence emerges as a good general strategy. In this section, we consider three qualitatively different environmental constraints favouring stable goals.

Resource rationality

In many cases, goal stability is likely to reflect efficient use of limited cognitive resources (Figure 1A) [19,20]. This follows from a resource rational perspective: humans often face a trade-off between maximising reward and reducing cognitive costs [24–29]. Across a range of settings, humans can rationally balance reward against costs, including the number of action sequences explored or the complexity of task information cognitively represented [28,30–34]. Structuring decision making around goals may optimise cognitive resource allocation while resulting in behaviours appearing rigid or overly persistent [20].

Consider a simple example – what to make for dinner. Let's say in the morning you might select the goal of making a lasagne. This initial goal selection can reduce later cognitive costs in multiple ways [20]. First, rather than pay attention to all information available in the environment (possible restaurants, or ingredients for other recipes), early goal selection allows you to focus on the subset of information most relevant for the current goal. This is consistent with recent work showing that people form efficient 'task construals' during planning [34,35]. By saving on the costs of representing irrelevant information, we are likely to see behaviours manifesting as over-commitment simply because alternative courses of action are processed less. This is supported by a recent study where we found that individual differences in goal commitment (in the face of better alternatives) correlated with ignoring irrelevant information: individuals with higher commitment were worse at reporting perceptual information about alternative goals as illustrated in Figure 2C [17].

Second, a substantial body of research has shown that switching between mental representations is itself cognitively costly [36–42] – for example, switching between representing the steps involved in making lasagne and those involved in making a curry. Goal persistence biases may, in part, reflect an avoidance of the cognitive costs associated with switching between dissimilar representations [19]. Notably, costs of switching representations can only be part of the picture in explaining goal commitment, because people often show striking flexibility in how they reconfigure their implementation of a goal, alongside rigidity in re-evaluating the goal itself [43]. For example, if the usual grocery store has run out of lasagne sheets, we might choose the costly alternative of driving all the way to a different shop (i.e., changing the implementation) rather than consider a different recipe (i.e., changing the goal).

Third, in complex environments, planning overall future states to evaluate action is often intractable [28]. Given these constraints, it may be resource rational to select goals using simpler heuristics and using more intensive forms of planning to implement them [43]. For example, we might pick a recipe first before planning exactly which shops to visit and how to get there. In this

case, goal stability could emerge from these sequential stages of selection and pursuit – an idea formalised in the 'Rubicon' family of models discussed later.

Finally, in social settings where plans are made with other agents, the costs of goal abandonment will quickly multiply [16,44,45]. Stable goals are critical to social coordination – committing to your plan of making lasagne enables you to ask your housemate to pick up the pasta sheets on their side of town, ensuring everyone's time and effort are used as efficiently as possible. In such cases, abandoning a goal carries both individual and collective costs, including potential reputational costs undermining future collaboration [44].

Shielding from interference

Aside from saving cognitive resources, goal commitment may stem from another, qualitatively different, adaptive function: shielding decisions from unwanted interference (Figure 1B). Consider the decision to cook dinner oneself rather than grab an over-priced and unhealthy take-out from the chain next door. While the right decision might feel clear in advance, it is easy to imagine making the opposite choice when passing the take-out place later that evening. This reversal in preferences is part of a well-studied phenomenon called 'hyperbolic discounting' [46–48]: people are able to forgo smaller rewards for later larger ones when both options are distant, but preferences are transiently reversed when the smaller reward is more imminent. Here, we suggest that mechanisms orienting us toward selected goals could reduce the impact of transient motivational shifts by shielding from goal-irrelevant options. For example, having set the goal of making lasagne, we might be less likely to notice the take-out restaurant as we walk past it (a shift in goal-directed attention), and even less likely to desire it (enhanced valuation of the chosen goal).

Philosophers have described this idea – the notion that setting prior intentions can be beneficial because they prevent us from revisiting decisions in moments of compromised or inferior decision making – as rational non-reconsideration [49–51]. While this intuition has been proposed by philosophers, empirical evidence that goals could actually suppress temptation comes from the literature on self-regulation [52–55]. For example, one study reported that priming a particular personal goal (e.g., 'studying') reduced the speed at which a relevant temptation word (e.g., 'basketball') was identified [52]. While this remains a relatively unexplored area within computational cognitive science, it raises important questions about the algorithms that might enable stable control in the face of fluctuating desires. A promising avenue for future research concerns whether individuals strategically select goals during periods of enhanced decision-making capacity, for example, under conditions of maximal access to information or cognitive resources, or minimal threat from short-term impulses.

Scaffolding motivation

In the previous sections, we emphasised that goal selection can benefit agents by constraining the option space – enabling more efficient processing or filtering interference. As a consequence, goal stability would arise simply because alternative options are no longer entertained. However, goals can also play a generative role in motivating action, expanding the option space. Humans have the capacity to be gripped by highly idiosyncratic goals, even ones with no obvious utility. This is perhaps most striking in the example of play [12]. Children commit to seemingly arbitrary goals – stacking blocks to a specific height or pursuing a self-imposed target in a game. Anyone who has observed children playing will know how absorbing these goals become, and how much frustration they elicit when impeded. These self-selected goals appear to generate their own reward signals, where the process of progressing toward the goal state becomes intrinsically satisfying, and abandonment becomes aversive. Mechanisms that generate intrinsic motivation toward selected goals would also result in the markers of goal commitment. In this case, we

would observe stable goals because of an added 'bonus' on goal-congruent options, rather than a filter on non-goal options.

In these cases, selecting a goal motivates us to act in the absence of external reward, sometimes even when the final goal state is known to be inconsequential. For example, having spent half an hour building a tower of blocks, a child might happily knock it over as soon as the last block is placed. Nevertheless, having these goals could be adaptive for a number of reasons, including facilitating exploration or practising new skills [12]. In this case, having these goals can offer long-term benefits (Figure 1C), while giving rise to what may appear to be 'irrational' attachment to the particular goal.

Scaffolding motivation through self-imposed goals does not disappear in adulthood. Instead, we propose that it remains particularly critical in contexts where the link between action and value requires additional support. In many cases, we choose goals based on dimensions of value that are too complex or abstract to easily and directly guide moment-to-moment action. The philosopher Henry Sidgwick illustrates this idea with the example of playing a sport [56] – for example, a game of football. He observes that the motivations for playing the game in the first place (e.g., spending time with friends or getting exercise) are different from the motivations driving behaviour once the game is underway. Once the game is in action, a more immediate motivation takes over: the urge to win. Here, aiming for a concrete goal state (to put the ball in the opponent's goal) directs action in a way that would be difficult to compute by deferring to the original values (to be healthy, to have friends, to have fun). In other words, once a goal is adopted for any number of complex value attributes, the goal state itself becomes the primary source of motivation. However, this can lead to behaviours that seem like over-commitment – persisting in pursuit of the goal even when those original values are no longer being served.

Unlike other animals, humans can sustain behaviour toward all manner of abstract value, sometimes at odds with basic needs (consider Saint Simeon Stylites, who chose to spend 36 years living on a pillar in the desert for the sake of spiritual purity). We propose that goals in this context enable us to generate basic reward signals toward attaining targets that align with these abstract values. This function of goals becomes particularly important when the source of value is too distant, intangible, or multiply realisable to effectively guide moment-to-moment behaviour. In this sense, abstract longer-term values can be 'scaffolded' by concrete goals eliciting more basic reward signals that direct action, while giving rise to the hallmarks of goal commitment.

The algorithmic level: implementing goal stability

What algorithms could support flexible behaviour while delivering these different benefits associated with goal stability? In this section, we examine several algorithmic architectures (Figure 3) with respect to the teleological considerations of efficient planning, shielding from temptation, and scaffolding motivation.

Algorithms separating goal selection and implementation

An extreme case of goal stability arises in a family of models under the umbrella of 'Rubicon models'. Crossing the Rubicon River – passing the point of no return – captures the idea that we are locked into selected goals. At the heart of these models is the premise that real-time decision making unfolds sequentially in discrete phases corresponding first to the evaluation of possible goals ('goal setting'), followed by pursuit of the selected goal ('goal implementation') [9,57–59]. Importantly, in these models, once a goal is chosen, deliberation over alternative goals ceases and resources are allocated to its implementation. It is clear why this separation of decision phases would manifest behaviourally in strong goal stability: if the goal itself is not subject to re-assessment, people will perseverate even in the face of strong evidence favouring its

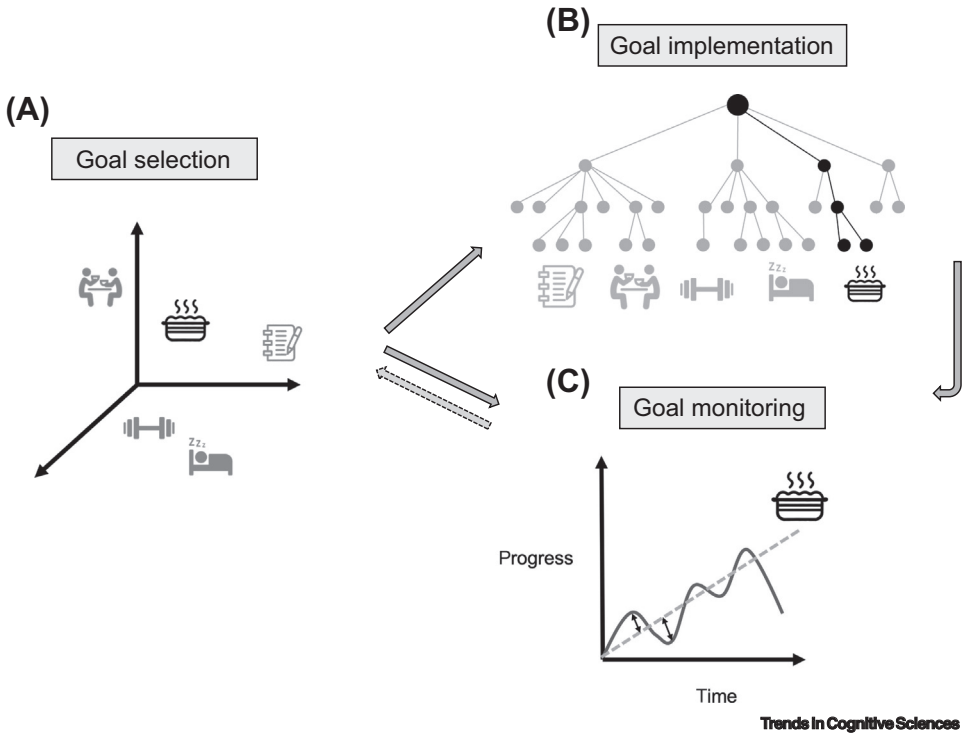

Figure 3. Algorithmic implementations of goal selection and pursuit. (A) Goal selection: different possible goals (e.g., making dinner, seeing friends, or finishing a paper) are plotted within a multi-dimensional value space, reflecting trade-offs across various attributes – for example, knowledge acquisition or social connection. Such high-level values may be difficult to compute on the fly and less effective at sustaining motivation without support from concrete, proximal goals providing internal reward signals during pursuit. In turn, these internal reward signals may act as a 'bonus' for the current goal, leading to greater goal stability. (B) Goal implementation: once a goal is selected, cognitive resources may shift from deliberation to implementation [9]. The tree-search diagram illustrates how only the pathways leading to the selected goal remain actively considered during planning, while alternatives are pruned from further evaluation [35]. This could serve the function of both optimising cognitive resources and shielding from interference, aligning with Rubicon models of decision making where goal selection and implementation operate in distinct phases [59]. (C) Goal monitoring: effective goal pursuit requires maintaining flexibility in case the current goal is unattainable or becomes increasingly costly. We propose that goal monitoring may rely on dedicated decision mechanisms evaluating the current goal's success rather than continuously re-evaluating alternatives, consistent with recent evidence that goal monitoring is biased toward assessing the current goal [17]. Monitoring may involve comparing the goal's value against a cached threshold, such as a long-term average or an expected rate of progress (illustrated in the plot with a dotted line), and involve specialised neural architectures that support decisions to disengage [63,66,80]. This family of algorithms could minimise cognitive load and interference while maintaining flexibility to changes in the attainability of the chosen goal.

abandonment. For that same reason, Rubicon models could also protect earlier decisions from undesirable forms of interference.

The logic of separating goal selection from implementation echoes principles from hierarchical reinforcement learning (Box 2). However, the discrete phases could correspond to a range of different algorithmic implementations. There is some evidence that people use different types of algorithms for goal selection and implementation in a manner consistent with resource rational strategies [43]. Given the computational demands of performing exhaustive forward planning overall goals, one possibility is to use approximate strategies for selecting a goal and then allocate more extensive planning resources to its implementation. For example, agents might select outcomes as goals that have tended to be rewarding in the past (e.g., cooking a Nigella Lawson recipe) using cached value estimates, while implementing more flexible model-based planning to

Box 2. Hierarchical reinforcement learning

Many of the computational mechanisms we describe – goal selection, implementation, and revision – align naturally with ideas from hierarchical reinforcement learning (HRL). HRL extends standard reinforcement learning by introducing temporally extended actions that allow agents to break down tasks into reusable subroutines. In one influential formulation, these subroutines are referred to as 'options' – sequences of actions bundled together [83–85]. For example, making tea can be broken down into the options of boiling the kettle and pouring water – each of which could be reused in other contexts (e.g., making coffee). Each option is defined by an initiation set (states where it can begin), a policy (actions taken), and a termination set (states where it ends). The structure of goals could be operationalised in this framework: selecting a goal corresponding to selecting an option, the means of goal implementation corresponding to the option policy, and the final goal state corresponding to the termination set. However, one important deviation between some formulations of HRL and our characterisation of goal pursuit lies in the flexibility of goal implementation. Whereas options in HRL are often fixed, learned routines, empirical work suggests that humans pursue goals using flexible, model-based planning (even when the goals themselves are stable). For instance, one study found that participants tended to use habitual strategies to select a goal (reducing computational demands), but used more flexible planning to implement the goal [43] – a pattern that could not be captured by formulations of HRL where option execution is cached in the initial selection. Capturing this flexibility requires an agent to perform model-based planning toward an option, where options guide the agent's subsequent search among candidate implementations.

pursue a selected goal (e.g., planning the shopping trip and following the recipe). Indeed, one study found that in a sequential decision-making task, participants tended to use model-free strategies for goal selection alongside model-based strategies to achieve the chosen goal [43]. By dividing selection and implementation in this way, people reduced the costs of evaluating the entire state space, while retaining flexible planning toward a preselected goal (Figure 3B).

Rubicon models offer one proposal for maintaining goal stability: goal selection inhibits deliberation over other goals and promotes planning toward the selected goal. The glaring challenge for Rubicon models is then to provide a good account of goal abandonment. When decision making unfolds in sequential phases of selection and pursuit, how do agents ever escape goals?

Algorithms supporting flexibility

What algorithms could enable the flexibility to abandon failing goals, while continuing to bestow the benefits associated with stable goals? One family of algorithms that balances these demands involves monitoring the current default option against a threshold for its re-assessment. For example, this may involve monitoring success with the current goal against a long-run average or an expected rate of progress – rather than continuously monitoring all the alternative options (Figure 3C). This type of algorithm and signatures of its neural implementation are found in many ecological decision-making settings [60–66]. For example, animals' decisions about when to leave a current foraging patch and explore elsewhere are effectively modelled by an algorithm that compares the value of the current patch with a stored estimate of the environmental average [2,60,63]. Recent work has proposed that valuation of current goals is influenced by the notion of 'momentum' – conceptualised as the product of progress and rate of progress, echoing the Newtonian formulation of momentum as the product of mass and velocity [18]. Crucially, continued monitoring of the current goal value – however formalised – is consistent with the suppression of goal-irrelevant information that prevents distraction from alternative goals. This mechanism aligns with empirical findings showing that goal abandonment is disproportionately driven by devaluation of the current goal relative to an equivalent increase in the value of alternative goals during ongoing pursuit (Figure 2D) [17]. These algorithms monitoring current goals could exist alongside those supporting their implementation, enabling flexibility in a manner that is both resource efficient and shields from interference.

If there are mechanisms dedicated to monitoring revision of ongoing goals, how well are they tuned to the different computational demands arising in commitment? Earlier, we discussed a

number of reasons why stable goals stem from adaptive mechanisms in ecological environments. These suggest that the degree of goal stability should depend on global variables. In the case of resource rationality, commitment should be highest in environments with multiple alternatives or where information-seeking is costly. In the case of protecting goals from interference, knowledge that the environment is likely to elicit temptation or distraction should favour greater commitment. Finally, if goals serve to scaffold motivation, we may see stronger markers of commitment to selected goals predicated on abstract dimensions of value (e.g., climbing a mountain) compared with those driven by basic reward signals (e.g., satisfying hunger). Future research could explore the extent to which revision mechanisms are calibrated to these varying meta-cognitive demands (see Outstanding questions).

More broadly, given the stability of goal pursuit once initiated, it is critical that agents learn to select goals that merit commitment in the long term. A full discussion of goal selection lies beyond the scope of this article, but the broader literature on intrinsic motivation offers a useful framework for considering the variables at play (see [67–69] for reviews). Algorithms supporting goal selection in open-ended environments have also been a topic of interest in autotelic reinforcement learning (Box 3). Here, we highlight an important complement to goal revision mechanisms: the need for goal selection algorithms that anticipate the subsequent stability of those chosen goals.

Goal stability and mental health

The computational perspective we adopted provides a useful framework for understanding the different algorithms underpinning goal selection, pursuit, and revision. Mistuning of these mechanisms may lead to divergent forms of over- or under-commitment to goals. An exciting avenue for future research is to examine how the calibration of these different mechanisms could map onto transdiagnostic dimensions of mental health.

Much of this review has focused on how goal setting constrains the space of options considered – both shielding goals from interference and allowing resources to be allocated toward their implementation. This requires mechanisms that direct attention toward goal-congruent information while suppressing irrelevant information [17]. Attentional filtering must be tuned to match task demands [70]: too little filtering leaves individuals vulnerable to distraction, while excessive filtering may block potentially important signals. Individual differences in how these filtering mechanisms are calibrated are likely to underlie specific forms of reduced goal commitment in mental health. For example, symptoms of attention deficit hyperactivity disorder (ADHD) involve increased distractibility associated with changes in top-down attention [71], suggesting a general reduction

Box 3. Autotelic reinforcement learning and goal selection

Autotelic reinforcement learning refers to a family of algorithms in which agents learn to generate their own goals in order to facilitate the acquisition of useful skills or knowledge, without relying on externally specified reward functions [86,87]. These systems are designed to support open-ended learning in complex or sparse environments by enabling agents to perform self-directed exploration without extrinsic rewards. Complementing this theoretical framework, recent empirical work has begun to investigate how humans select goals in open-ended environments [88,89]. In one study, participants were asked to propose playful goals in a virtual reality setting of a child's bedroom (e.g., the goal of 'throwing a ball into the bin'). Participants tended to propose goals that contained reusable components [89], consistent with other work showing humans prefer plans with components that are not only simple but reusable [90]. More broadly, the values that underpin human goal selection are often highly abstract. We have suggested that intrinsic reward signals generated by progress toward a selected goal may be critical to our capacity to work toward these abstract values. For example, the high-level values underpinning goals such as climbing a mountain or learning the violin are remarkably distant from our basic motivational drives, and using these values to guide moment-to-moment action may be difficult to compute on the fly. In these cases, mechanisms that generate internal rewards as an agent progresses toward a concrete selected goal state could serve as motivational scaffolds for sustaining pursuit.

in shielding attention from goal-irrelevant information. By contrast, trait anxiety has been linked to more selective difficulties in suppressing threat-related information during goal pursuit [72]. Future work could investigate how mechanisms shielding ongoing goals from interference relate to these dimensions of mental health.

We have also highlighted the need for algorithms that support monitoring and revision of ongoing goals. An intriguing recent connection has been made between trait apathy and failures to monitor and revise current behaviour [73]. While apathy is a transdiagnostic symptom associated with general loss of motivation [74], it may not always manifest as failures to act, but rather failures to initiate a change in ongoing behaviour. In a recent sequential decision paradigm, trait apathy was paradoxically associated with over-persisting with an action sequence, marked by a failure to disengage at the right moment [73]. These findings suggest at least some apathy metrics relate to processes that monitor ongoing goals, rather than purely deficits in motivation. More broadly, while over-persistence and failures to persist with goals appear in opposition, these findings illustrate how both behaviours could occur in the same individuals if goal regulation is compromised. This is consistent with studies showing that trait apathy and impulsivity – traditionally viewed as opposites – are highly correlated in young adults [72], consistent with more general difficulties tuning goal engagement to the appropriate factors. An exciting future direction would be to capture these bidirectional failures computationally, in settings where goal monitoring is noisy, biased, or compromised.

Finally, regulation of the motivational signals that promote pursuit of goals and deter abandonment will be essential for maintaining mental wellbeing [75]. Goal pursuit is associated with a rich taxonomy of emotions, including frustration or dejection when goals are impeded and elation when goals are met [76]. Since failure to progress with goals is aversive, there has long been interest in the adaptive value of 'goal disengagement' – the idea that when goals become unattainable, individuals must be capable of detaching from this system of motivational signals to protect their wellbeing [76–78]. In a meta-analysis, goal disengagement (i.e., the capacity to detach from goals) was reliably associated with better wellbeing – including lower stress and negative affect – in the general population [79]. However, in individuals at risk of depression, goal disengagement was paradoxically associated with higher depressive symptoms, possibly because depressive symptoms themselves facilitate over-disengaging from goals. These findings highlight our need to regulate the signals motivating pursuit of chosen goals and the relevance of this capacity to mental health.

Concluding remarks

Humans have the extraordinary capacity to be captivated by the goals we choose. This often manifests as a strong reluctance to abandon goals – even in the face of high costs or better alternatives. Rather than viewing this tendency as a maladaptive bias, we suggest it reflects an adaptive set of mechanisms that solve key computational challenges. Stable goals may stem from optimisation of cognitive resources, shielding long-term goals from interference, and scaffolding complex behaviour in the absence of basic reward signals. Considering these adaptive functions will be key to developing algorithmic models capturing the central role of goals in decision making. In turn, characterising the different ways these processes could be miscalibrated will provide crucial insights for understanding disruptions of goal commitment affecting mental health.

Outstanding questions

Can people dynamically adjust their commitment to goals based on relevant environmental factors?

What computational roles do emotional signals play in shaping decisions during goal pursuit – emotions such as frustration, boredom, dejection, regret, anticipation, and elation?

What are the different 'failure modes' of goal commitment, and how do the mechanisms involved relate to distinct dimensions of mental health?

How do environmental and developmental experiences shape individual differences in goal commitment across the lifespan?

What are the evolutionary origins of commitment, and to what extent is goal stability uniquely human or shared with other species?

How do decision mechanisms supporting goal pursuit relate to model-free habitual behaviour?

When and how do goals become intrinsically rewarding, generating their own motivational signals independent of external reward?

How do goals interact across different levels of hierarchy, and how does commitment to a higher-level goal relate to commitment to sub-goals?

How could the algorithms supporting goal selection and pursuit in biological brains be implemented in current artificial learning systems, for example, in autotelic agents?

What mechanisms enable switching between multiple active goals in dynamic environments?

Acknowledgments

Artwork for Figure 1 was created by Magdalena Adomeit. We thank Branson Byers, Matan Mazor, and Richard Holton for helpful discussions and comments on the manuscript. E.H. was supported by a C.V. Starr Fellowship at Princeton Neuroscience Institute.

Declaration of interests

The authors declare no competing interests.

References

1. Arkes, H.R. and Blumer, C. (1985) The psychology of sunk cost. *Organ. Behav. Hum. Decis. Process.* 35, 124–140
2. Hayden, B.Y. *et al.* (2011) Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.* 14, 933–939
3. Blanchard, T.C. and Hayden, B.Y. (2015) Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS One* 10, e0117057
4. Constantino, S.M. and Daw, N.D. (2015) Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.* 15, 837–853
5. Sugawara, M. and Katahira, K. (2021) Dissociation between asymmetric value updating and perseverance in human reinforcement learning. *Sci. Rep.* 11, 3574
6. Eckstein, M.K. *et al.* (2022) The interpretation of computational model parameters depends on the context. Hartley C, Behrens TE, Radulescu A, editors. *eLife* 11, e75474
7. Marr, D. (2010) *Vision: a computational investigation into the human representation and processing of visual information [Internet]*, The MIT Press [cited 2025 Apr 17]. Available from: https://direct.mit.edu/books/monograph/3299/VisionA-Computational-Investigation-into-the-Human
8. Juechems, K. and Summerfield, C. (2019) Where does value come from? *Trends Cogn. Sci.* 23, 836–850
9. O'Reilly, R.C. (2020) Unraveling the mysteries of motivation. *Trends Cogn. Sci.* 24, 425–434
10. Molinaro, G. and Collins, A.G.E. (2023) A goal-centric outlook on learning. *Trends Cogn. Sci.* 27, 1150–1164
11. De Martino, B. and Cortese, A. (2023) Goals, usefulness and abstraction in value-based choice. *Trends Cogn. Sci.* 27, 65–80
12. Chu, J. *et al.* (2024) In praise of folly: flexible goals and human cognition. *Trends Cogn. Sci.* 28, 628–642
13. Austin, J.T. and Vancouver, J.B. (1996) Goal constructs in psychology: structure, process, and content. *Psychol. Bull.* 120, 338–375
14. Kruglanski, A.W. (1996) Goals as knowledge structures. In *The Psychology of Action: Linking Cognition and Motivation to Behavior*, pp. 599–618, The Guilford Press, New York, NY, US
15. Chu, J. and Schulz, L. (2022) 'Because I want to': valuing goals for their own sake. *Proc. Ann. Meet. Cogn. Sci. So c. [Internet]* 44 [cited 2024 Mar 18]. Available from: https://escholarship.org/uc/item/6dg5w6cf
16. Cheng, S. *et al.* (2023) Intention beyond desire: spontaneous intentional commitment regulates conflicting desires. *Cognition* 238, 105513
17. Holton, E. *et al.* (2024) Goal commitment is supported by vmPFC through selective attention. *Nat. Hum. Behav.* 1–15
18. Aenugu, S. and O'Doherty, J.P. (2025) Building momentum: a computational account of persistence toward long-term goals. *PLoS Comput. Biol.* 21, e1013054
19. Molinaro, G. *et al.* (2025) *Investigating the role of representation switching costs in goal persistence bias*, Second Workshop on Representational Alignment https://openreview.net/forum?id=odTK4tps2C
20. Prystawski, B. *et al.* (2022) Resource-rational models of human goal pursuit. *Top. Cogn. Sci.* 14, 528–549
21. Staw, B.M. (1976) Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action. *Organ. Behav. Hum. Perform.* 16, 27–44
22. Kahneman, D. and Tversky, A. (1979) Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263
23. Thaler, R. (1980) Toward a positive theory of consumer choice. *J. Econ. Behav. Organ.* 1, 39–60
24. Simon, H.A. (1955) A behavioral model of rational choie. *Q. J. Econ.* 69, 99–118
25. Anderson, J.R. (2013) *The Adaptive Character of Thought*, Psychology Press, New York
26. Gershman, S.J. *et al.* (2015) Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* 349, 273–278
27. Griffiths, T.L. *et al.* (2015) Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. *Top. Cogn. Sci.* 7, 217–229
28. Callaway, F. *et al.* (2022) Rational use of cognitive resources in human planning. *Nat. Hum. Behav.* 6, 1112–1125
29. Lieder, F. and Griffiths, T.L. (2020) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43, e1
30. Huys, Q.J.M. *et al.* (2015) Interplay of approximate planning strategies. *Proc. Natl. Acad. Sci.* 112, 3098–3103
31. Sezener, C.E. *et al.* (2019) Optimizing the depth and the direction of prospective planning using information values. *PLoS Comput. Biol.* 15, e1006827
32. Correa, C.G. *et al.* (2023) Humans decompose tasks by trading off utility and computational cost. *PLoS Comput. Biol.* 19, e1011087
33. Lai, L. and Gershman, S.J. (2024) Human decision making balances reward maximization and policy compression. *PLoS Comput. Biol.* 20, e1012057
34. Ho, M.K. *et al.* (2023) Rational simplification and rigidity in human planning. *Psychol. Sci.* 34, 1281–1292
35. Ho, M.K. *et al.* (2022) People construct simplified mental representations to plan. *Nature* 606, 129–136
36. Allport, D.A. *et al.* (1994) Shifting intentional set: exploring the dynamic control of tasks. In *Attention and Performance 15: Conscious and Nonconscious Information Processing*, pp. 421–452, The MIT Press, Cambridge, MA, US
37. Rogers, R.D. and Monsell, S. (1995) Costs of a predictable switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* 124, 207–231
38. Monsell, S. (2003) Task switching. *Trends Cogn. Sci.* 7, 134–140
39. Kiesel, A. *et al.* (2010) Control and interference in task switching–a review. *Psychol. Bull.* 136, 849–874
40. Koch, I. *et al.* (2018) Cognitive structure, flexibility, and plasticity in human multitasking-An integrative review of dual-task and task-switching research. *Psychol. Bull.* 144, 557–583
41. Musslick, S. *et al.* (2017) Multitasking Capability Versus Learning Efficiency in Neural Network Architectures. *Proc. 39th Annu. Meet. Cogn. Sci. Soc. London* 829–834
42. Grahek, I. *et al.* (2025) Control adjustment costs limit goal flexibility: empirical evidence and a theoretical account. *Psychol. Rev.*
43. Cushman, F. and Morris, A. (2015) Habitual control of goal selection in humans. *Proc. Natl. Acad. Sci.* 112, 13817–13822
44. Intention, Bratman M. (1987) *Plans, and Practical Reason*, Harvard University Press, Cambridge: Cambridge, MA
45. Bahrami, B. *et al.* (2010) Optimally interacting minds. *Science* 329, 1081–1085
46. Ainslie, G. and Haslam, N. (1992) Hyperbolic discounting. In *Choice Over Time*, pp. 57–92, Russell Sage Foundation, New York, NY, US
47. Loewenstein, G. and Prelec, D. (1992) Anomalies in intertemporal choice: evidence and an interpretation. *Q. J. Econ.* 107, 573–597
48. Laibson, D. (1997) Golden eggs and hyperbolic discounting. *Q. J. Econ.* 112, 443–477
49. Holton, R. (1999) Intention and weakness of will. *J. Philos.* 96, 241
50. Broome, J., ed (2013) *Rationality Through Reasoning*, Malden, MA, Wiley-Blackwell
51. Bratman, M. (2014) Temptation and the agent's standpoint. *Inquiry* 57, 293–310
52. Fishbach, A. *et al.* (2003) Leading us not into temptation: momentary allurements elicit overriding goal activation. *J. Pers. Soc. Psychol.* 84, 296–309
53. Fujita, K. and Sasota, J.A. (2011) The effects of construal levels on asymmetric temptation-goal cognitive associations. *Soc. Cogn.* 29, 125–146
54. Milyavskaya, M. *et al.* (2015) Saying "no" to temptation: want-to motivation improves self-regulation by reducing temptation rather

than by increasing self-control. *J. Pers. Soc. Psychol.* 109, 677–693
55. Fishbach, A. and Hofmann, W. (2015) Nudging self-control: a smartphone intervention of temptation anticipation and goal resolution improves everyday goal progress. *Motiv. Sci.* 1, 137–150
56. Sidgwick, H. (1907) *The Methods of Ethics [Internet]* (7th ed.), Macmillan, p. 71 [cited 2025 Apr 17]. Book IV. Available from: https://www.gutenberg.org/ebooks/46743
57. Klinger, E. (1975) Consequences of commitment to and disengagement from incentives. *Psychol. Rev.* 82, 1–25
58. Heckhausen, H. and Gollwitzer, P.M. (1987) Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motiv. Emot.* 11, 101–120
59. Achtziger, A. and Gollwitzer, P. (2008) Motivation and volition in the course of action. In *Motivation and Action* (2010) (2nd ed) (Heckhausen, J., ed.), pp. 275–299, Cambridge University Press, New York
60. Stephens, D.W. and Krebs, J.R. (1986) *Foraging Theory*, Princeton University Press
61. Charnov, E.L. (1976) Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.* 9, 129–136
62. Cisek, P. (2012) Making decisions through a distributed consensus. *Curr. Opin. Neurobiol.* 22, 927–936
63. Kolling, N. *et al.* (2012) Neural mechanisms of foraging. *Science* 336, 95–98
64. Hayden, B.Y. and Moreno-Bote, R. (2018) A neuronal theory of sequential economic choice. *Brain Neurosci. Adv.*, 2398212818766675
65. Hayden, B.Y. (2018) Economic choice: the foraging perspective. *Curr. Opin. Behav. Sci.* 24, 1–6
66. Ahmadlou, M. *et al.* (2025) A subcortical switchboard for perseverative, exploratory and disengaged states. *Nature* 1–11
67. Ryan, R.M. and Deci, E.L. (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78
68. Oudeyer, P.Y. *et al.* (2007) Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286
69. Gottlieb, J. *et al.* (2013) Information seeking, curiosity and attention: computational and neural mechanisms. *Trends Cogn. Sci.* 17, 585–593
70. Lavie, N. (2005) Distracted and confused?: selective attention under load. *Trends Cogn. Sci.* 9, 75–82
71. Friedman-Hill, S.R. *et al.* (2010) What does distractibility in ADHD reveal about mechanisms for top-down attentional control? *Cognition* 115, 93–103
72. Sharp, P.B. *et al.* (2022) Humans perseverate on punishment avoidance goals in multigoal reinforcement learning. Gillan CM, *et al*, editors. *eLife* 11, e74402
73. Scholl, J. *et al.* (2022) The effect of apathy and compulsivity on planning and stopping in sequential decision-making. *PLoS Biol.* 20, e3001566
74. Heron, C.L. *et al.* (2019) Brain mechanisms underlying apathy. *J. Neurol. Neurosurg. Psychiatry* 90, 302–312
75. Johnson, S.L. *et al.* (2010) Goal dysregulation in the affective disorders. In *Emotion Regulation and Psychopathology: A Transdiagnostic Approach to Etiology and Treatment*, pp. 204–228, The Guilford Press, New York, NY, US
76. Carver, C.S. and Scheier, M.F. (1998) *On the Self-Regulation of Behavior*, Cambridge University Press, New York, NY, US, p. xx
77. Wrosch, C. *et al.* (2003) The importance of goal disengagement in adaptive self-regulation: when giving up is beneficial. *Self Identity* 2, 1–20
78. Wrosch, C. *et al.* (2003) Adaptive self-regulation of unattainable goals: goal disengagement, goal reengagement, and subjective well-being. *Personal. Soc. Psychol. Bull.* 29, 1494–1508
79. Barlow, M.A. *et al.* (2020) Goal adjustment capacities and quality of life: a meta-analytic review. *J. Pers.* 88, 307–323
80. Tervo, D.G.R. *et al.* (2021) The anterior cingulate cortex directs exploration of alternative strategies. *Neuron* 109, 1876–1887.e6
81. Velez-Ginorio, J. *et al.* (2017) Interpreting actions by attributing compositional desires. *Proc. Ann. Meet. Cogn. Sci. Soc. [Internet]* 39 [cited 2025 Feb 25]. Available fro https://escholarship.org/uc/item/3qw110xj
82. Jara-Ettinger, J. *et al.* (2020) The Naïve Utility Calculus as a unified, quantitative framework for action understanding. *Cogn. Psychol.* 123, 101334
83. Sutton, R.S. *et al.* (1999) Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artif. Intell.* 112, 181–211
84. Precup, D. (2000) *Temporal abstraction in reinforcement learning*, University of Massachusetts Amherst
85. Botvinick, M.M. *et al.* (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113, 262–280
86. Colas, C. *et al.* (2019) Curious: intrinsically motivated modular multi-goal reinforcement learning. In *International Conference on Machine Learning [Internet]*, pp. 1331–1340, PMLR
87. Colas, C. *et al.* (2022) Autotelic agents with intrinsically motivated goal-conditioned reinforcement le http://arxiv.org/abs/2012.09830
88. Molinaro, G. *et al.* (2024) Latent Learning Progress Drives Autonomous Goal Selection in Human Reinforcement Learning. In *NeurIPS 2024-The Thirty-Eighth Annual Conference on Neural Information Processing Systems [Internet]*
89. Davidson, G. *et al.* (2025) Goals as reward-producing programs. et al. 7, 205–220
90. Correa, C.G. *et al.* (2025) Exploring the hierarchical structure of human plans via program generation. *Cognition*, 105990