Learning telic-controllable state representations

Nadav Amir (nadav.amir@princeton.edu)
Princeton Neuroscience Institute, Princeton University
Princeton, NJ 08540 USA

Stas Tiomkin (stas.tiomkin@ttu.edu)
Department of Computer Science, Texas Tech University
Lubbock, TX 79409 USA

Abstract

Computational models of purposeful behavior comprise both descriptive and prescriptive aspects, used respectively to ascertain and evaluate situations in the world. In reinforcement learning, prescriptive reward functions are assumed to depend on predefined and fixed descriptive state representations. Alternatively, these two aspects may emerge interdependently: goals can shape the acquired state representations and vice versa. Here, we present a computational framework for state representation learning in bounded agents, where descriptive and prescriptive aspects are coupled through the notion of goal-directed, or telic, states. We introduce the concept of telic-controllability to characterize the tradeoff between the granularity of a telic state representation and the policy complexity required to reach all telic states. We propose an algorithm for learning telic-controllable state representations, illustrating it using a simulated navigation task. Our framework highlights the role of deliberate ignorance – knowing what to ignore – for learning state representations that balance goal flexibility and cognitive complexity. **Keywords:** State representation learning; Controllability; Rate-Distortion Theory

Figure 1: **The granularity-complexity tradeoff:** within the framework developed in this paper, state representations partition all experiences (blue ellipses) into preference-ordered classes called "telic states" (curved regions), each consisting of all experience distributions that are approximately equivalent with respect to the agent's goal. **Left:** an agent with a coarse-grained goal is unable to reach a desired telic state, S, which is too distant, in a statistical sense, from its default policy π_0. **Right:** refining the goals yields a state representation that is more controllable since all telic states can be reached using incremental policy update steps.

Introduction

How do goals shape the way learning agents represent their experience? This fundamental question has only recently started drawing increased attention in both cognitive science and AI (Molinaro & Collins, 2023; Muhle-Karbe et al., 2023; Radulescu, Niv, & Ballard, 2019; Eysenbach, Zhang, Levine, & Salakhutdinov, 2022; Florensa, Held, Geng, & Abbeel, 2018; M. Wang, Jin, & Montana, 2024). For example, recent empirical work suggest that humans may structure their experience into discrete states by balancing the utility and complexity of their representations (Fang & Sims, 2025). However, it remains unclear, from a theoretical perspective, how should computationally bounded learning agents adjust their state representations when their goals are changing. For example, consider a rodent navigating a complex maze with changing reward contingencies (Krausz et al., 2023), or a robot trained to do various object manipulation tasks with sparse rewards (Andrychowicz et al., 2017). How should such learning agents represent their tasks in ways that facilitate adaptation to shifting goals using limited computational resources? Prior works have addressed the problem of efficient state representation learning using bisimulation (Zhang, McAllister, Calandra, Gal, & Levine, 2020; Z. Wang, Wang, Xiao, Zhu, & Stone, 2024) or option-based methods (Abel et al., 2020). While these approaches can provide efficient heuristics for state abstraction in Markovian settings, our aim here is to characterize the fundamental tradeoff between the granularity of a state representation (which may or may not be Markovian), and the computational resources needed to generate a policy that can efficiently utilize it. We present a principled approach to this problem, leveraging a recently proposed theoretical framework of goal-directed, or *telic*, state representation learning (Amir, Niv, & Langdon, 2023). We define a novel property, *telic-controllability*, characterizing the ability to reach all states within a given telic state representation using complexity bound policies. We describe a telic state representation learning algorithm and illustrate it using a simple navigation task by showing how complexity bounded agents can learn a telic-controllable state representation that can adapt to shifting goals.

Formal setting

Telic states as goal-equivalent experiences

We assume the setting of a perception-action cycle, i.e., sequences of observation-action pairs representing the flow of information between agent and environment. We denote by O and $\mathcal{A}$ the set of possible observations and actions, respectively. An experience sequence, or *experience* for short, is a finite sequence of observation-action pairs: $h = o_1, a_1, o_2, a_2, ..., o_n, a_n$. For every non-negative integer, $n \geq 0$, we denote by $\mathcal{H}_n \equiv (O \times \mathcal{A})^n$ the set of all experiences of length n. The collection of all finite experiences is denoted

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

by $\mathcal{H} = \cup_{n=1}^{\infty} \mathcal{H}_n$. In non-deterministic settings, it will be useful to consider distributions over experiences rather than individual experiences themselves and we denote the set of all probability distributions over finite experiences by $\Delta(\mathcal{H})$. Following Bowling et al. (2022), we define a *goal* as a binary preference relation over experience distributions. For any pair of experience distributions, $A, B \in \Delta(\mathcal{H})$, we write $A \succeq_g B$ to indicate that experience distribution A is weakly preferred by the agent over B, i.e., that A is at least as desirable as B, with respect to goal g. When $A \succeq_g B$ and $B \succeq_g A$ both hold, A and B are equally preferred with respect to g, denoted as $A \sim_g B$. We observe that $\sim_g$ is an equivalence relation, i.e., it satisfies the following three properties, for any $A, B, C \in \Delta(\mathcal{H})$:

- Reflexivity: $A \sim_g A$ for all $A \in \Delta(\mathcal{H})$.
- Symmetry: $A \sim_g B$ implies $B \sim_g A$ for all $A, B \in \Delta(\mathcal{H})$.
- Transitivity: if $A \sim_g B$ and $B \sim_g C$ then $A \sim_g C$ for all $A, B, C \in \Delta(\mathcal{H})$.

Therefore, every goal induces a partition of $\Delta(\mathcal{H})$ into disjoint sets of equally desirable experience distributions. For goal g, we define the goal-directed, or *telic*, state representation, $\mathcal{S}_g$, as the partition of experience distributions into equivalence classes it induces: $\mathcal{S}_g = \Delta(\mathcal{H}) / \sim_g$. In other words, each telic state represents a generalization over all equally desirable experience distributions. This definition captures the intuition that agents need not distinguish between experiences that are equivalent, in a statistical sense, with respect to their goal. Furthermore, since different telic states are, by definition, non-equivalent with respect to $\succeq_g$, the goal g also determines whether a transition between any two telic states brings the agent in closer alignment to, or further away from its goal.

Learning with telic states

How can telic state representations guide goal-directed behavior? To address this question, we recall the definition of a *policy*, π, as a distribution over actions given the past experience sequence and current observation:

$$\pi(a_i | o_1, a_1, ..., o_i). \tag{1}$$

Analogously, we can define an *environment*, e, as a distribution over observations given the past experience sequence:

$$e(o_i | o_1, a_1, ..., a_{i-1}). \tag{2}$$

The distribution over experience sequences can be factored, using the chain rule, as follows:

$$P_\pi(o_1, a_1, ..., o_n, a_n) = P(o_1, a_1, ..., o_n, a_n | e, \pi) = \prod_{i=1}^{n} e(o_i | o_1, a_1, ..., a_{i-1}) \pi(a_i | o_1, a_1, ..., o_i). \tag{3}$$

Typically, the environment is assumed to be fixed, and hence not explicitly parameterized in $P_\pi(h)$ above. The definition of telic states as goal-induced equivalence classes can now be extended to equivalence classes of policy-induced experience distributions as follows:

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \tag{4}$$

The question we are interested in here is the following: how can an agent learn an efficient policy for reaching a desired telic state? In other words, how can an agent acquire policies that are likely to generate experiences belonging to a certain telic state $S_i \in \mathcal{S}_g$? To answer this, we consider the empirical distribution of N experience sequences generated by policy π:

$$\hat{P}_\pi(h) = \frac{|\{i : h_i = h\}|}{N}. \tag{5}$$

By Sanov's theorem (Cover, 1999), the likelihood that $\hat{P}_\pi(h)$ belongs to telic state S_i decays exponentially with a rate of

$$R = D_{KL}(P_i^\star || P_\pi) \tag{6}$$

where,

$$P_i^\star = arg \min_{P \in S_i} D_{KL}(P || P_\pi), \tag{7}$$

is the *information projection* of P_π onto S_i, i.e., the distribution in S_i which is closest, in the KL sense, to P_π. Thus, R can be thought of as the "telic distance" from π to S_i since it determines the likelihood that experiences sampled from P_π belong to the telic state S_i. Assuming a policy parameterized by θ, the following policy gradient method updates π_θ in a way that minimizes its telic distance to S_i:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta D_{KL}(P_i^\star || P_{\pi_\theta}), \tag{8}$$

where $\eta > 0$ is a learning rate parameter.

Experience features and discrimination sensitivity

A natural way of representing goals, i.e., preferences over experience distributions, is by comparing the likelihood that experiences generated from different distributions will belong to some subset $\Phi_g \subset \mathcal{H}$ representing some desired property of experiences. For example, for the goal of solving a maze, Φ_g might be the set of all experiences, i.e., path trajectories, that reach the exit. Formally, for two experience distributions, A and B, the agent will prefer the one that is more likely to generate experiences belonging to Φ_g:

$$A \succeq_g B : \sum_{h \in \Phi_g} A(h) \geq \sum_{h \in \Phi_g} B(h).$$

The sensitivity parameter, ε, effectively determining the maximum difference, in terms of desirable outcome likelihoods, that the agent is willing to ignore in order to reduce representational complexity. In the maze example, experience distribution A would be preferred over B if it is more likely to generate trajectories that reach the exit. Importantly, Eq. implies that A and B are equivalent only when $\sum_{h \in \Phi_g} A(h)$ and $\sum_{h \in \Phi_g} B(h)$ are precisely equal, which is unlikely in realistic,

noisy environments. A more reasonable assumption is that agents can discriminate sampling likelihoods at some finite sensitivity level, $\varepsilon > 0$, such that:

$$A \sim_g^{(\varepsilon)} B \iff |\sum_{h\in\Phi_g} A(h) - \sum_{h\in\Phi_g} B(h)| \leq \varepsilon. \quad (9)$$

In the maze example, this means that two trajectory distributions are considered equivalent if their respective likelihoods of generating exit-reaching trajectories are within ε of each other. As we shall see in the following sections, the discrimination sensitivity parameter, ε, controls the tradeoff between the granularity of a telic state representation and the policy complexity needed to reach all telic states.

Telic-controllability

In this section, we introduce the notion of *telic-controllability*, a joint property of an agent and a telic state representation, that characterizes whether or not the agent is able to reach all possible telic states using complexity-limited policy update steps. Towards this, we first define an agent's *policy*, π, as a distribution over actions given the past experience sequence and current observation: $\pi(a_i|o_1,a_1,...,o_i)$. Assuming a fixed environment, the definition of telic states as goal-induced equivalence classes induces corresponding equivalence classes of policy-induced experience distributions as follows:

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \quad (10)$$

As detailed above, this mapping between policies and telic states provides a unified account of goal-directed learning in terms of the statistical distance between policy-induced distributions and desired telic states. To explore this notion, we introduce a new property – telic-controllability – that plays a central role in the following sections. A representation is called telic-controllable if any state can be reached using a finite number, N, of complexity-limited policy updates, starting from the agent's default policy, π_0, where the complexity of a policy update step is quantified by the Kullback-Leibler (KL) divergence between the post and pre-update step policies. Formally, we have the following:

Definition (telic-controllability)**.** A telic-state representation, $\mathcal{S}_g$, induced by the goal, g, is *telic-controllable* with respect to a default policy, π_0, and a policy complexity capacity, $\delta \geq 0$, if the following holds:

$$\begin{aligned} &\forall S \in \mathcal{S}_g \; \exists \{\pi_t, S_t\}_{t=0}^{N}, N > 0 \text{ s.t. } \forall t < N \\ &\big(S_t = [P_{\pi_t}]_{\sim_g}\big) \wedge \big(D_{KL}(P_{\pi_{t+1}}||P_{\pi_t}) \leq \delta\big) \wedge \big(S_N = S\big), \end{aligned} \quad (11)$$

where $[P_{\pi_t}]_{\sim_g}$ is the goal-induced equivalence class, i.e., telic state, containing P_{π_t}. This definition generalizes the familiar control theoretic notion of controllability in two important ways. First, it applies to telic states, i.e., classes of distributions over action-outcome trajectories, rather than by n-dimensional vectors – the standard control theoretic setting. Second, it takes into account the complexity capacity limitations of the agent, using information theoretic quantifiers to constrain the maximal complexity of policy update steps an agent can take in attempting to reach one telic state from another. As illustrated in the next section, telic-controllability is a desirable property since it means that agents can flexibly adjust to shifting goals using bounded policy complexity resources.

State representation learning algorithm

A central feature of our approach is the duality it establishes between goals and state representations. In this section, we utilize this duality to develop an algorithm for learning a telic-controllable state representation, or, equivalently, finding a goal that produces such a state representation. The algorithm receives as inputs the agent's current goal, g (represented, e.g., by an ordered set of desired experience features), and default policy, π_0, along with its policy complexity capacity, δ, and the discrimination sensitivity parameter ε. Its output consists of a new goal g' such that $\mathcal{S}_{g'}$ is telic controllable with respect to π_0 and δ. The main idea is to split any unreachable telic state, S, i.e., one that cannot be reached from π_0 using policy update steps with complexity less than δ. State splitting is accomplished by generating a new, intermediate, telic state, S_M, lying between the agent's default policy induced distribution, P_{π_0}, and its information projection on the unreachable telic state, i.e., the distribution $P^* \in S$ that is closest to P_{π_0}, in the KL sense. The intermediate telic state, S_M, is then defined as the set of all distributions that are ε-equivalent to P_M (Eq. 9), where P_M is the convex combination of P^* and P_{π_0} lying at a KL distance of δ from P_{π_0}. After generating the new state, S_M, the goal is updated to reflect the proper ordering between the default policy state S_0, the intermediate state S_M, and the originally unreachable state S, such that elements of S_M are between S_0 and S in terms of preference. Pseudocode for the learning algorithm is provided in Algorithm 1. The algorithm makes use of an auxiliary procedure, FINDREACHABLESTATES (Algorithm 2), to find all reachable states, given the agent's goal, g, default policy, π_0, and policy complexity constraint, δ. This auxiliary procedure performs a recursive search, similar to depth-first search methods, attempting to find policies that are closest, in the KL sense, to currently unreachable telic states, while still sufficiently close to the agent's current policy, as not to exceed the policy complexity capacity. Its main optimization step (line 3) can be implemented, e.g., using policy gradient over the information projection of P_{π_0} on S.

Illustrative example: dual goal navigation task

In this section, we illustrate the proposed state representation framework and learning algorithm using a simple navigation task in which an agent performs a one dimensional random walk, starting at location $x_0 = 0$, with the goal of reaching one of two non overlapping regions of interest after a fixed number, $T = 30$, of steps. The agent's policy is defined as a stochastic mapping between its current and next position and

Algorithm 1 Telic-controllable state representation learning

Input: π_0: default policy, g: current goal,
δ: policy complexity capacity, ε: sensitivity.
Output: g': new goal such that $\mathcal{S}_{g'}$ is telic-controllable with respect to π_0 and δ

$\mathcal{R} \leftarrow [P_{\pi_0}]_{\sim_g}$ ▷ initialize reachable state set
$g' \leftarrow g$ ▷ initialize new goal
while $\mathcal{R} \neq \mathcal{S}_{g'}$ **do**
  $\mathcal{R} \leftarrow$ FINDREACHABLESTATES(π_0, g', δ) ▷ see algorithm 2 below
  for $S \in \mathcal{S}_{g'} \setminus \mathcal{R}$ **do** ▷ for each unreachable state
    $P^* \leftarrow \arg\min_{P \in S} D_{KL}(P||P_{\pi_0})$ ▷ information projection of P_{π_0} on S
    $M = \arg\max_{t \in [0,1]} t$ s.t. $D_{KL}\big((tP^* + (1-t)P_{\pi_0})||P_{\pi_0}\big) \leq \delta$
    $P_M = MP^* + (1-M)P_{\pi_0}$ ▷ convex combination of P^* and P_{π_0}
    $S_M \leftarrow \{P : P \sim_g^{(\varepsilon)} P_M\}$ ▷ ε-neighborhood of P_M
    if $P_{\pi_0} \leq_g P^*$ **then** ▷ update goal with preference order for S_M
      $g' \leftarrow g' \cup \{(p,q)_{\leq_{g'}} \in S_M \times S\} \cup \{(r,p)_{\leq_{g'}} \in S_0 \times S_M\}$
    else if $P^* \leq_g P_{\pi_0}$ **then**
      $g' \leftarrow g' \cup \{(q,p)_{\leq_{g'}} \in S \times S_M\} \cup \{(p,r)_{\leq_{g'}} \in S_M \times S_0\}$
    end if
  end for
end while
return g'

Algorithm 2 Finding reachable states

Input: π_0: initial policy, g: goal,
δ: policy complexity constraint.
Output: all telic states in $\mathcal{S}_g$ reachable from π_0 by δ-complexity limited policy update steps

procedure RECURSIVEREACH$(\pi, g, \delta, \mathcal{R})$
  for $S \in \mathcal{S}_g \setminus \mathcal{R}$ **do** ▷ for every unreached state S
    $\pi_\theta \leftarrow \arg\min_\theta D_{KL}(S||P_{\pi_\theta})$ s.t. $D_{KL}(P_{\pi_\theta}||P_\pi) \leq \delta$ ▷ optimize policy to reach S
    if $[P_{\pi_\theta}]_{\sim_g} \notin \mathcal{R}$ **then** ▷ if new state reached
      $\mathcal{R} \leftarrow \mathcal{R} \cup [P_\pi]_{\sim_g}$ ▷ add current state to reachable set
      $\mathcal{R} \leftarrow$ RECURSIVEREACH$(\pi_\theta, g, \delta, \mathcal{R})$ ▷ continue from current state
    end if
  end for
  return $\mathcal{R}$
end procedure
procedure FINDREACHABLESTATES(π_0, g, δ)
  $\mathcal{R}_0 \leftarrow [P_{\pi_0}]_{\sim_g}$ ▷ initialize reachable set
  $\mathcal{R} \leftarrow$ RECURSIVEREACH$(\pi_0, g, \delta, \mathcal{R}_0)$ ▷ try to reach all states recursively
  return $\mathcal{R}$ ▷ return set of reachable states
end procedure

is parameterized by the mean and standard deviation (μ and σ, respectively) of a Gaussian update step: $\pi(x_{t+1}|x_t;\mu,\sigma) = x_t + \eta_t, \quad \eta_t \sim \mathcal{N}(\mu,\sigma)$. For brevity, we denote by $\pi(\mu,\sigma)$ a policy with a $\mathcal{N}(\mu,\sigma)$ distributed noise term. A graphical illustration of the task and sample trajectories for different policies is shown in Fig. 2.

Since the sum of normally distributed variables is also normally distributed, a policy $\pi(\mu,\sigma)$ induces a Gaussian distribution over the final location of the agent:

$$p(x_T \mid x_0 = 0;\mu,\sigma) = \mathcal{N}(T\mu, \sqrt{T}\sigma). \tag{12}$$

To account for goal-directed behavior, we define a right and a left region of interest, R and L, consisting of unit radius segments centered around $x_R = 2$ and $x_L = -2$ respectively. Thus, $R = [R_1, R_2] = [1,3]$ and $L = [L_1, L_2] = [-3,-1]$. For the purpose of this example, we assume that the agent wants to reach R but avoid L, at time T. For example, for a rodent navigating a narrow corridor, R and L may indicate segments of the corridor where a reward (e.g., food) and a punishment (e.g., air puff) are administered, respectively. We can express the agent's goal in terms of preferences over policies by defining $\Delta P(\mu,\sigma) = p(x_T \in R \mid \mu,\sigma) - p(x_T \in L \mid \mu,\sigma)$ as the difference between the probabilities that the agent will reach regions R and L at time T, with a policy $\pi(\mu,\sigma)$. The agent's goal can now be defined as a preference for policies with higher ΔP values. However, as explained above, due to the agent's finite discrimination resolution, it can only detect whether ΔP is above or below the sensitivity threshold, ε. Thus, using Eq. 10, the agent's goal, g, can be expressed by the following preference relation over policies, where we denote, for brevity, $\pi(\mu_i,\sigma_i)$ and $\Delta P(\mu_i,\sigma_i)$ as π_i and ΔP_i, respectively, for $i = 1,2$:

$$\pi_1 \succeq_g \pi_2 \iff \big(\Delta P_1 \geq \varepsilon \geq \Delta P_2\big) \vee \big(\Delta P_1 \geq -\varepsilon \geq \Delta P_2\big), \tag{13}$$

where first term on the r.h.s. of Eq. 13 captures the *desirability* of R – the agent prefers policies that have a probability *higher* than ε of reaching R over ones that do not; while the second term captures the *undesirability* of L – the agent prefers policies that have a probability *lower* than ε to reach L than ones that do not. We recall that telic states can be defined by policies that are similarly preferred, under the agent's discrimination threshold, ε, which determines the borders between the resulting telic states. The telic state representation for the goal g defined by Eq. 13, and a threshold parameter of $\varepsilon = 0.1$ is visualized in Fig. 3 (top left). Telic state S_R (S_L), is shown as a colored region bounded by a dotted green (red) line, consisting of all policies that are more (less) likely to reach R than L by a probability margin of ε or more. Policies that are roughly equally likely to reach R or L, i.e., whose difference in ΔP is smaller than ε, constitute an additional "default" telic state, S_0 (teal background), in which the agent

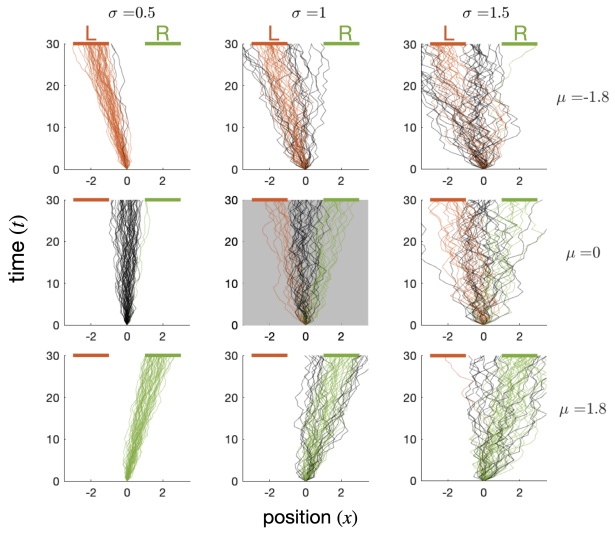

Figure 2: **Dual goal navigation task:** each tile shows 500 one-dimensional random walk trajectories of length $T = 30$, generated by a Gaussian policy parameterized by the mean (μ) and standard deviation (σ) of position update step (x-axis) across time (y-axis). Regions of interest R and L consist of line segments centered around $x_R = 2$ and $x_L = -2$, shown as green and red lines respectively at $T = 30$. Trajectories reaching one of the goals are plotted in the corresponding color, illustrating the relationship between policy parameters and goal reaching likelihoods. The default policy, $(\mu_0, \sigma_0) = (0, 1)$, shown in the center gray tile, is equally likely to reach R and L.

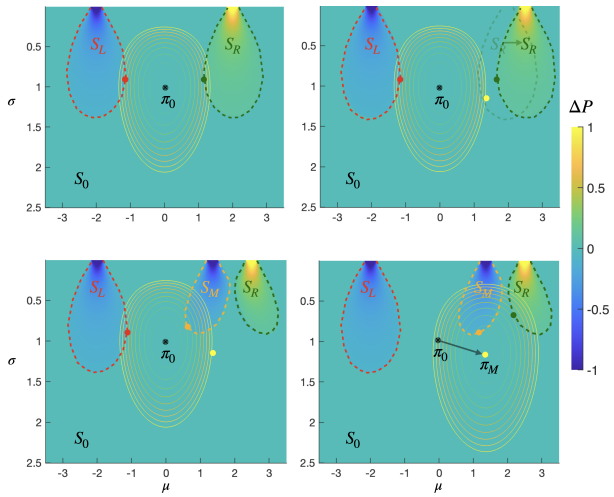

Figure 3: **Telic state representation learning for navigation task with shifting goals:** points in (μ, σ) policy space colored by the difference between their probability of reaching unit length regions, R and L, centered around 2 and -2 respectively, at time $T = 30$. **Top left:** telic states S_L and S_R (outlined by red and green dashed lines, respectively) consist of policies that are more likely to reach the corresponding region by a threshold of $\varepsilon = 0.1$ or more. Contour lines indicate isometric policy complexity levels, relative to the default policy $\pi_0 : (\mu_0 = 0, \sigma_0 = 1)$ (black dot), for a capacity bound of $\delta = 1$ bit. Green and red dots show the information projection of π_0 on S_R and S_L respectively, i.e., the policies each telic state closest to π_0 in KL-divergence **Top right:** shifting the center of R to 2.5, renders S_R unreachable from π_0 with δ bounded policy complexity. The policy $\pi_M : (\mu_M, \sigma_M)$ (yellow dot) is the one closest to S_R while still within the complexity capacity of the agent. **Bottom left:** splitting S_R by inserting an intermediate telic-state, S_M, centered around μ_M. By construction, the nearest distribution to π_0 in S_M, in the KL sense (orange dot), is within the agent's complexity capacity. **Bottom right:** both S_M and S_R are reachable with respect to the agent's new default policy, $\pi_M(\mu_M = 1.37, \sigma_M = 1.15)$ (see algorithm 1 for details); the new telic state representation $\{S_0, S_L, S_M, S_R\}$ is telic controllable with respect to $\pi_0(0, 1)$, $\delta = 1$, and $N = 1$.

is agnostic to which region is it more likely to reach.

$$
\begin{aligned}
S_R &= \{(\mu,\sigma) | \Delta P(\mu,\sigma) \geq \varepsilon\}, \\
S_L &= \{(\mu,\sigma) | \Delta P(\mu,\sigma) \leq -\varepsilon\}, \\
S_0 &= \{(\mu,\sigma) | |\Delta P(\mu,\sigma)| \leq \varepsilon\}.
\end{aligned} \tag{14}
$$

Using Eqs. 12 and 14 we can express each telic state in closed form, for example S_R can be expressed, using the standard error function, $\text{erf}(x) = 2/\sqrt{\pi} \int_0^x e^{-t^2} dt$, as follows:

$$
\begin{aligned}
S_R = \{(\mu,\sigma) \Big| &\frac{1}{2}\left(\text{erf}\frac{R_1 - T\mu}{\sqrt{2T}\sigma} - \text{erf}\frac{R_2 - T\mu}{\sqrt{2T}\sigma}\right) - \\
&\frac{1}{2}\left(\text{erf}\frac{L_1 - T\mu}{\sqrt{2T}\sigma} - \text{erf}\frac{L_2 - T\mu}{\sqrt{2T}\sigma}\right) \geq \varepsilon\},
\end{aligned}
$$

with similar expressions for S_L and S_0. To illustrate the notion of telic-controllability (Eq. 11) using this representation, we define the complexity, $C(\pi)$, of a policy, $\pi(\mu, \sigma)$, with respect to the agent's default policy, $\pi_0(\mu_0, \sigma_0)$, as the KL divergence, per time step, between them:

$$
C(\pi) \equiv D_{KL}(\pi \| \pi_0).
$$

The contour lines in the first three panels of Fig. 3 (top & bottom left) show isometric policy complexity levels for an agent with a complexity capacity of $\delta = 1$ bit per time step, and a default policy $\pi_0(\mu_0 = 0, \sigma_0 = 1)$. Initially, both telic states, S_R, and S_L, lie within the range of the agent's policy complexity capacity (top left). The policies in S_R and S_L that are closest in the KL sense to π_0 (green and red dots, respectively), both lie within a range of less than δ from π_0, i.e., the state representation is telic-controllable. When the center of R shifts from $x_R = 2$ to $x_R = 2.5$ (top right), telic state S_R is no longer within complexity range δ from π_0 and the state representation becomes non-controllable. To address this (bottom left), the state representation learning algorithm described in , splits S_R by adding an intermediate telic state S_M (orange), centered around the policy closest to S_R that is still within a KL-range of δ from π_0 (yellow dot). This changes the shape of S_R and S_L since now the probability of reaching each of the three telic states, S_R, S_L and S_M, is defined in with respect to the two others, e.g., $S_M = \{(\mu,\sigma) | \Delta P_M(\mu,\sigma) \geq \varepsilon\}$ where $\Delta P_M = p(x_T \in M \mid \mu,\sigma) - \max\{p(x_T \in L \mid \mu,\sigma), p(x_T \in R \mid \mu,\sigma)\}$, and similarly for S_R and S_L. Since π_M is, by construction, within a KL range of δ from π_0, the agent can reach S_M by updating its default policy to π_M (bottom right), bringing S_R into reach again. Hence, the new state representation, consisting of S_0, S_L, S_M and S_R, is tellic-controllable. Fig. 5 illustrates the telic-complexity curves, showing the probability

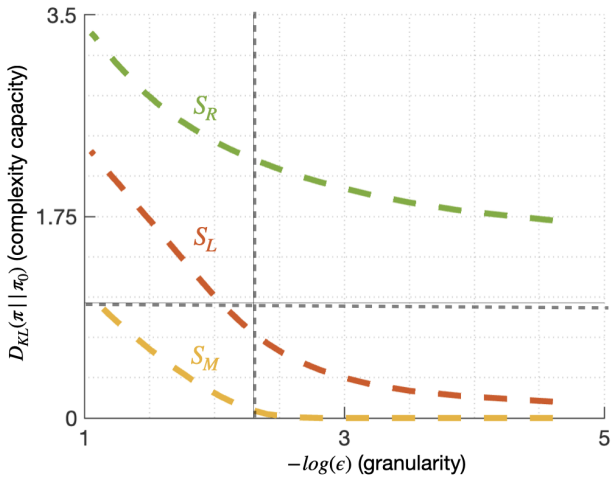

Figure 4: **Complexity-granularity curves:** Each line shows the policy complexity capacity, relative to the default policy $\pi_0(0,1)$ (ordinate) required to reach the corresponding telic state at a given representational granularity level, quantified by the negative log of the sensitivity parameter ε (abscissa). Dashed gray lines show the values used in the dual-goal navigation example: $\delta = 1$ (horizontal) and $\varepsilon = 0.1$ (vertical)

of reaching each telic state achievable for a given complexity capacity level (x-axis). These curves quantify the maximal gain in the probability of reaching each telic state, S_R, S_M or S_L, relative to the other two (ordinate), for a given policy complexity capacity level, with respect to a default policy of π_0 (left) or π_M (right) (abscissa). Finally, Fig. 4 illustrates the granularity-complexity tradeoff: the granularity of the state representation, quantified as $-\log(\varepsilon)$ (abscissa), controls the complexity capacity required to reach each state (ordinate). Finer-grained representations are generally more controllable. For a granularity level of $\varepsilon = 0.1$ (gray vertical line), only S_L and S_M are reachable from $\pi_0(0,1)$ under a complexity capacity of $\delta = 1$ (gray horizontal line).

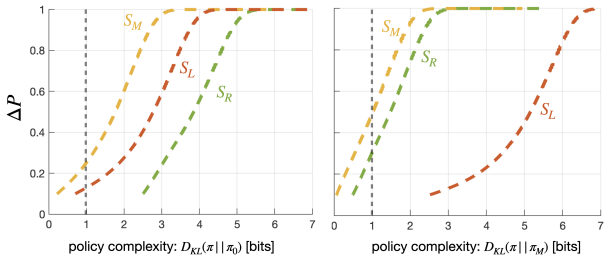

Figure 5: **Goal-complexity tradeoff curves:** the probability of reaching each telic state as a function of policy complexity. **Left:** an agent with a default policy $\pi_0 : (\mu_0, \sigma_0) = (0,1)$ is unable to reach telic state S_R with a complexity capacity limit of $\delta = 1$ (gray vertical line). **Right:** with $\pi_1 : (\mu_1, \sigma_1) = (1.09, 1.24)$ as its default policy, the agent can reach both S_M and S_R with the same policy complexity capacity.

Discussion

We illustrated a novel approach to modeling purposeful behavior in bounded agents, based on the hypothesis that goals, defined as preferences over experience distributions, play a fundamental role in shaping state representations. Coupling together descriptive and normative aspects of learning models, our framing posits a granularity-complexity tradeoff as a normative theoretical criterion guiding cognitive agents in determining which features of their environment to attend to and which to ignore in the context of a particular task (Niv et al., 2015; Langdon, Song, & Niv, 2019). We accordingly hypothesize that goal selection can be usefully viewed as a processes of balancing two competing cognitive loads: representational granularity and policy complexity. The former limits the resolution of the goals (and the corresponding telic state representation) that the agent selects, while the latter controls the complexity of the policy generation, preventing behavior Clearly, our approach is highly simplified and entails theoretical assumptions which may do not hold in the general case. For example, computing the telic-distance, requires that the agent knows, or at least has a good model of the environment dynamics, which may not be the case for complex, real-world environments. While beyond the scope of this paper, these limitations could be potentially addressed using estimation and learning theoretic methods for bounding the telic-distance error under partially-observed approximations of the environment dynamics. While several methods for goal-directed state abstraction have been previously proposed (Li, Walsh, & Littman, 2006; Shah et al., 2021; Kaelbling, 1993; Zhang et al., 2020; Steccanella & Jonsson, 2022) our approach is different in suggesting that telic state representations are only defined with respect to a goal (rather than, for example, defining goals as a subset of preexisting states). Our approach is aligned with work using resource rational analysis to explain human learning and representation (Prystawski, Mohnert, Tošić, & Lieder, 2022; Lieder & Griffiths, 2020; Ho et al., 2022; Correa et al., 2025) but our emphasis here is on developing a principled theoretical account of how goals shape state representation learning in complexity constrained cognitive agents. Our quantification of policy complexity follows previous work applying information theoretic principles in reinforcement learning (Tishby & Polani, 2010; Rubin, Shamir, & Tishby, 2012) and cognitive science (Amir et al., 2020; Lai & Gershman, 2024). Notably, our complexity-granularity curves (Fig. 4) qualitatively resemble rate-distortion curves in information theory (Cover, 1999), suggesting a new interpretation of state representation learning via information theoretic lens (Arumugam & Van Roy, 2021; Abel et al., 2019). Finally, the duality between goals and state representations characterizing our approach may help address the thorny problem of goal formation: where do goals come from in the first place? Specifically, goals may be selected based on the properties of the state representations they produce. Our framework thus suggests that bounded agents, who need to balance control over the environment with behavioral adaptability (cf. Klyubin, Polani, and Nehaniv (2005)), would do well to choose goals that produce telic-controllable state representations.

Acknowledgments

This work was supported by grant no. U01DA050647 from the National Institute on Drug Abuse.

References

Abel, D., Arumugam, D., Asadi, K., Jinnai, Y., Littman, M. L., & Wong, L. L. (2019). State abstraction as compression in apprenticeship learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 3134–3142).

Abel, D., Umbanhowar, N., Khetarpal, K., Arumugam, D., Precup, D., & Littman, M. (2020). Value preserving state-action abstractions. In *International conference on artificial intelligence and statistics* (pp. 1639–1650).

Amir, N., Niv, Y., & Langdon, A. (2023). States as goal-directed concepts: an epistemic approach to state-representation learning. *Information-Theoretic Principles in Cognitive Systems Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:2312.02367*.

Amir, N., Suliman-Lavie, R., Tal, M., Shifman, S., Tishby, N., & Nelken, I. (2020). Value-complexity tradeoff explains mouse navigational learning. *PLOS Computational Biology*, *16*(12), e1008497.

Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., ... Zaremba, W. (2017). Hindsight experience replay. *Advances in neural information processing systems*, *30*.

Arumugam, D., & Van Roy, B. (2021). Deciding what to learn: A rate-distortion approach. In *International conference on machine learning* (pp. 373–382).

Bowling, M., Martin, J. D., Abel, D., & Dabney, W. (2022). Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*.

Correa, C. G., Sanborn, S., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2025). Exploring the hierarchical structure of human plans via program generation. *Cognition*, *255*, 105990.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Eysenbach, B., Zhang, T., Levine, S., & Salakhutdinov, R. R. (2022). Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 35603–35620.

Fang, Z., & Sims, C. R. (2025). Humans learn generalizable representations through efficient coding. *Nature Communications*, *16*(1), 3989.

Florensa, C., Held, D., Geng, X., & Abbeel, P. (2018). Automatic goal generation for reinforcement learning agents. In *International conference on machine learning* (pp. 1515–1528).

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*(7912), 129–136.

Kaelbling, L. P. (1993). Learning to achieve goals. In *Ijcai* (Vol. 2, pp. 1094–8).

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation* (Vol. 1, pp. 128–135).

Krausz, T. A., Comrie, A. E., Kahn, A. E., Frank, L. M., Daw, N. D., & Berke, J. D. (2023). Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron*, *111*(21), 3465–3478.

Lai, L., & Gershman, S. J. (2024). Human decision making balances reward maximization and policy compression. *PLOS Computational Biology*, *20*(4), e1012057.

Langdon, A. J., Song, M., & Niv, Y. (2019). Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making. *Behavioural Processes*, *167*, 103891.

Li, L., Walsh, T. J., & Littman, M. L. (2006). Towards a unified theory of state abstraction for MDPs. In *Ai&m*.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Molinaro, G., & Collins, A. G. E. (2023). A goal-centric outlook on learning. *Trends in Cognitive Sciences*.

Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W., & Summerfield, C. (2023). Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, *111*(23), 3885–3899.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157.

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

Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: the role of structure and attention. *Trends in cognitive sciences*, *23*(4), 278–292.

Rubin, J., Shamir, O., & Tishby, N. (2012). Trading value and information in MDPs. *Decision Making with Imperfect Decision Makers*, 57–74.

Shah, D., Xu, P., Lu, Y., Xiao, T., Toshev, A., Levine, S., & Ichter, B. (2021). Value function spaces: Skill-centric state abstractions for long-horizon reasoning. *arXiv preprint arXiv:2111.03189*.

Steccanella, L., & Jonsson, A. (2022). State representation learning for goal-conditioned reinforcement learning. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 84–99).

Tishby, N., & Polani, D. (2010). Information theory of decisions and actions. In *Perception-action cycle: Models, architectures, and hardware* (pp. 601–636). Springer.

Wang, M., Jin, Y., & Montana, G. (2024). Goal-conditioned offline reinforcement learning through state space partitioning. *Machine Learning*, 1–31.

Wang, Z., Wang, C., Xiao, X., Zhu, Y., & Stone, P. (2024). Building minimal and reusable causal state abstractions for reinforcement learning. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 38, pp. 15778–15786).

Zhang, A., McAllister, R., Calandra, R., Gal, Y., & Levine, S. (2020). Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*.