# Learning telic-controllable state representations

**Nadav Amir**
nadav.amir@princeton.edu
Princeton Neuroscience Institute
Princeton University

**Stas Tiomkin**
stas.tiomkin@sjsu.edu
Department of Computer Engineering
San Jose State University

**Angela Langdon**
angela.langdon@nih.gov
National Institute of Mental Health
National Institutes of Health

## Abstract

Computational descriptions of purposeful behavior comprise both *descriptive* and *normative* aspects. The former are used to ascertain current (or future) states of the world and the latter to evaluate the desirability, or lack thereof, of these states under some goal. In Reinforcement Learning, the normative aspect (reward and value functions) is assumed to depend on a predefined and fixed descriptive one (state representation). Alternatively, these two aspects may emerge interdependently: goals can be, and indeed often are, approximated by state-dependent reward functions, but they may also shape the acquired state representations themselves. Here, we present a novel computational framework for state representation learning in bounded agents, where descriptive and normative 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 controllable state representations, illustrating it using a simple navigation task with shifting goals. Our framework highlights the crucial role of deliberate ignorance – knowing which features of experience to ignore – for learning state representations that balance goal flexibility and policy complexity. More broadly, our work advances a unified theoretical perspective on goal-directed state representation learning in natural and artificial agents.


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 (roughly) 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 $\pi_0$. **Right:** by refining the goal the agent can acquire a more controllable state representation, where all telic states can be reached via simple policy update steps.

## 1 Introduction

How do goals shape the state representations acquired by learning agents? This question is the focus of growing attention in both cognitive science (Molinaro & Collins, 2023; Muhle-Karbe et al., 2023; Radulescu et al., 2019) and reinforcement learning (Eysenbach et al., 2022; Florensa et al., 2018; Wang et al., 2024a). A fundamental open problem however, is how to acquire useful state representations when goals are changing and computational resources are bounded. 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 environments 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 et al., 2020; Wang et al., 2024b) or option-based methods (Abel et al., 2020). While these approaches can provide efficient heuristics for state abstraction in Markovian settings, our focus here is on the fundamental relationship between the granularity of state representations, which may be non-Markovian, and the policy complexity resources needed to efficiently utilize them. We present a principled approach to this problem, leveraging a recently proposed theoretical framework of goal-directed, or *telic*, state representation learning (Amir et al., 2023). We define a novel property, *telic-controllability*, characterizing the ability of reaching 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.

## 2 Formal setting

### 2.1 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 $\mathcal{O}$ 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 $\mathcal{H}_n \equiv (\mathcal{O} \times \mathcal{A})^n$ the set of all experiences of length $n$. The collection of all finite experiences is denoted 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. See Amir et al. (2023) for additional details.

### 2.2 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 a goal of solving a maze, $\Phi_g$ 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 $\Phi_g$:

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

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. 2.2 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, $\epsilon > 0$, such that:

$$A \sim_g^{(\epsilon)} B \iff |\sum_{h \in \Phi_g} A(h) - \sum_{h \in \Phi_g} B(h)| \leq \epsilon. \tag{1}$$

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

### 2.3 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*, $\pi$, 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 can be extended to equivalence between policy-induced experience distributions as follows:

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \tag{2}$$

As detailed in appendix. A, this mapping between polices and telic states provides an unified account of goal-directed learning in terms 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, $\pi_0$, 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, $\pi_0$, and a policy complexity capacity, $\delta \geq 0$, if the following holds:

$$\forall S \in \mathcal{S}_g \; \exists \{\pi_t, S_t\}_{t=0}^N, N > 0 \text{ s.t. } \forall t < N \; \big(S_t = [P_{\pi_t}]_{\sim_g}\big) \wedge \big(D_{KL}(P_{\pi_{t+1}}||P_{\pi_t}) \leq \delta\big) \wedge \big(S_N = S\big), \tag{3}$$

where $[P_{\pi_t}]_{\sim_g}$ is the goal-induced equivalence class, i.e., telic state, containing $P_{\pi_t}$. 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 demonstrated in section 3 below, telic-controllability is a desirable property since it means that agents can flexibly adjust to shifting goals using bounded policy complexity resources.

### 2.4 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, $\pi_0$, along with its policy complexity capacity, $\delta$, and the discrimination sensitivity parameter $\epsilon$. It output consists of a new goal $g'$ such that $\mathcal{S}_{g'}$ is telic controllable with respect to $\pi_0$ and $\delta$. The main idea is to split any unreachable telic state, $S$, i.e., one that cannot be reached from $\pi_0$ using policy update steps with complexity less than $\delta$. State splitting is accomplished by generating a new, intermediate, telic state, $S_M$, lying between the agent's default policy induced distribution, $P_{\pi_0}$, and its information projection on the unreachable telic state, i.e., the distribution $P^* \in S$ that is closest to $P_{\pi_0}$, in the KL sense. The intermediate telic state, $S_M$, is then defined as the set of all distributions that are $\epsilon$-equivalent to $P_M$ (Eq. 1), where $P_M$ is the convex combination of $P^*$ and $P_{\pi_0}$ lying at a KL distance of $\delta$ from $P_{\pi_0}$. 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 appendix. B.

## 3 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 is parameterized by the mean and standard deviation ($\mu$ and $\sigma$, 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 a normally distributed, a policy $\pi(\mu, \sigma)$ induces a Gaussian distribution over the final location of the agent:

$$p(x_T \mid x_0 = 0; \mu, \sigma) = \mathcal{N}(T\mu, \sqrt{T}\sigma). \tag{4}$$

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 $\Delta P$ values. However, as explained in section 2.2 above, due to the agent's finite discrimination resolution, it can only detect

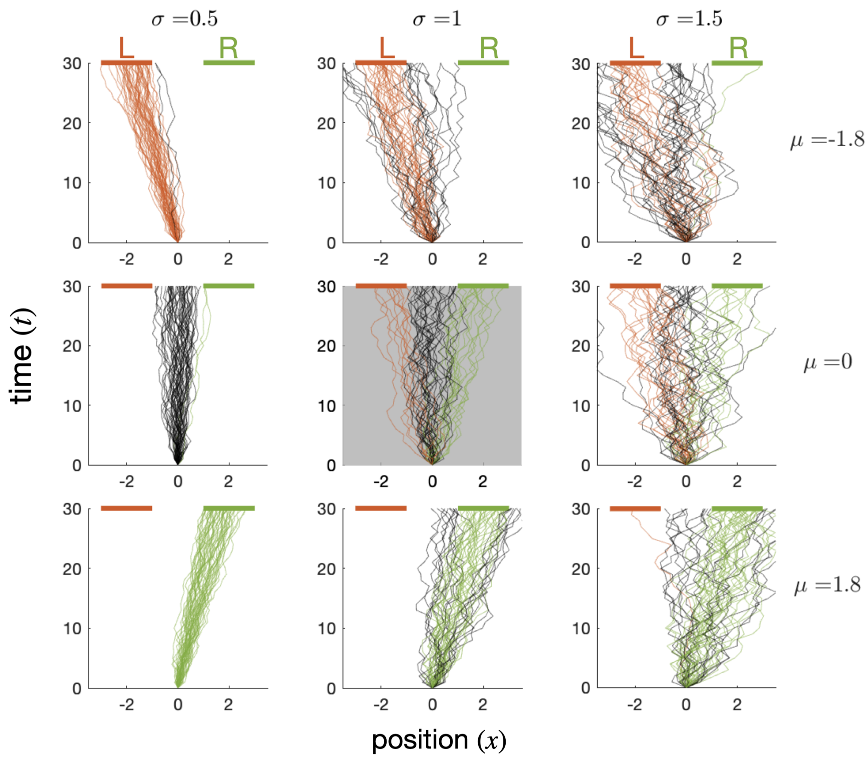


Figure 2: **Dual goal navigation task:** each tile shows 500 one-dimensional random walk trajectories of length $T = 30$, generated by a Gaussian policy parameterized by the mean ($\mu$) and standard deviation ($\sigma$) of position update step (x-axis) across time (y-axis). Regions of interest $R$ and $L$ 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 $L$.

whether $\Delta P$ is above or below the resolution threshold, $\epsilon$. Thus, using Eq. 2, the agent's goal, $g$, can be expressed by the following preference relation over policies:

$$\pi_1(\mu_1, \sigma_1) \succeq_g \pi_2(\mu_2, \sigma_2) \iff \big(\Delta P(\mu_1, \sigma_1) \geq \epsilon \geq \Delta P(\mu_2, \sigma_2)\big) \vee \big(\Delta P(\mu_1, \sigma_1) \geq -\epsilon \geq \Delta P(\mu_2, \sigma_2)\big), \tag{5}$$

where first term on the r.h.s. of Eq. 5 captures the *desirability* of $R$ – the agent prefers policies that have a probability *higher* than $\epsilon$ 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 $\epsilon$ to reach $L$ than ones that do not. The discrimination threshold, $\epsilon$, thus determines the borders between the resulting telic states. The telic state representation for the goal $g$ defined by Eq. 5, and a threshold parameter of $\epsilon = 0.1$ 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 $\epsilon$ or more. Policies that are roughly equally likely to reach $R$ or $L$, i.e., whose difference in $\Delta P$ is smaller than $\epsilon$, constitute an additional "default" telic state, $S_0$ (teal background), in which the agent is agnostic to which region is it more likely to reach.

$$S_R = \{(\mu, \sigma)|\Delta P(\mu, \sigma) \geq \epsilon\}, S_L = \{(\mu, \sigma)|\Delta P(\mu, \sigma) \leq -\epsilon\}, S_0 = \{(\mu, \sigma)||\Delta P(\mu, \sigma)| \leq \epsilon\}. \tag{6}$$

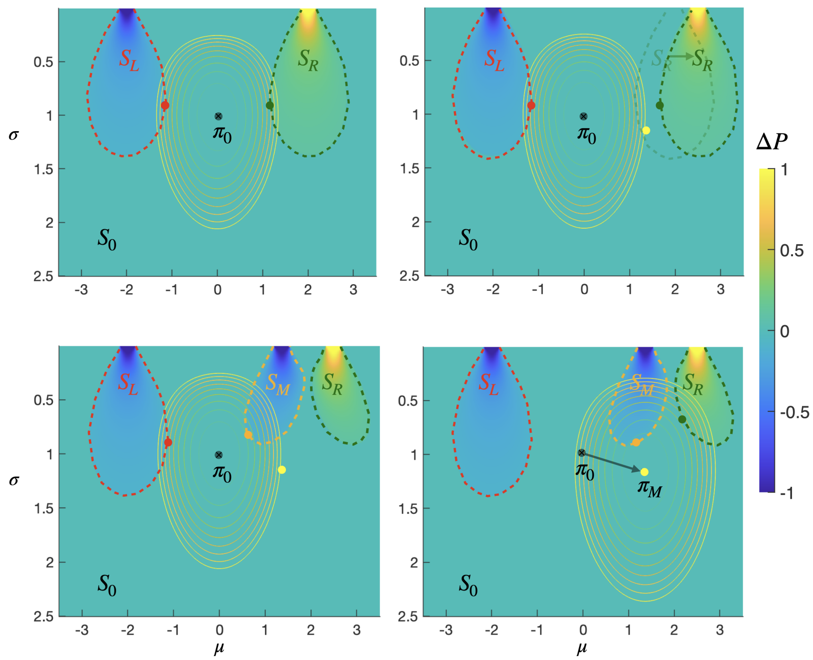


Figure 3: **Telic state representation learning for navigation task with shifting goals:** points in $(\mu, \sigma)$ policy space colored by the difference between their probability of reaching unit length regions, $R$ and $L$, centered around 2 and $-2$ respectively, at time $T = 30$. **Top left:** telic states $S_L$ and $S_R$ (blue and red dashed regions) consist of policies that are more likely to reach the corresponding region by a threshold of $\epsilon = 0.1$ 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 $\pi_0$ on $S_R$ and $S_L$ respectively, i.e., the policies each telic state closest to $\pi_0$ in KL-divergence **Top right:** shifting the center of $R$ to 2.5, renders $S_R$ unreachable from $\pi_0$ with $\delta$ 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 $\mu_M$. By construction, the nearest distribution to $\pi_0$ in $S_M$, in the KL sense (orange dot), is within the agents complexity capacity. **Bottom right:** both $S_M$ and $S_R$ are reachable with respect to the agent's new deafult 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 $N = 1$.

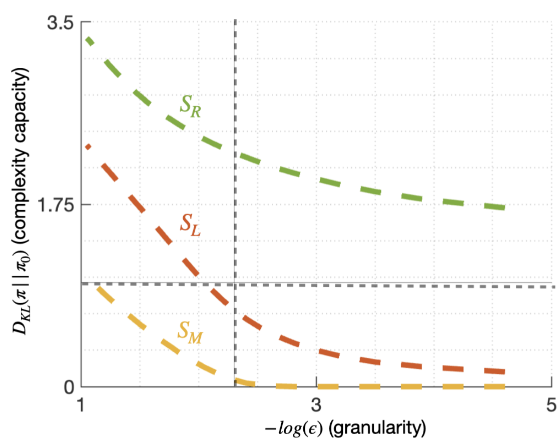


Figure 4: **Complexity-granularity curves:** Each line shows the policy complexity capacity, relative to the default policy $\pi_0(0, 1)$ (ordinate) required to reach the corresponding telic state at a given representational granularity level, quantified by the negative log of the sensitivity parameter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

Using Eqs. 4 and 6 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:

$$S_R = \left\{(\mu, \sigma)\middle|\frac{1}{2}\left(\text{erf}\,\frac{R_1 - T\mu}{\sqrt{2T}\sigma} - \text{erf}\,\frac{R_2 - T\mu}{\sqrt{2T}\sigma}\right) - \frac{1}{2}\left(\text{erf}\,\frac{L_1 - T\mu}{\sqrt{2T}\sigma} - \text{erf}\,\frac{L_2 - T\mu}{\sqrt{2T}\sigma}\right) \geq \epsilon\right\},$$

with similar expressions for $S_L$ and $S_0$. To illustrate the notion of telic-controllability (Eq. 3) 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 capcity (top left). The policies in $S_R$ and $S_L$ that are closest in the $KL$ sense to $\pi_0$ (green and red dots, respectively), both lie within a range of less than $\delta$ from $\pi_0$, 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 $\delta$ from $\pi_0$ and the state representation becomes non-controllable. To address this (bottom left), the state representation learning algorithm described in 2.4, 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 $\delta$ from $\pi_0$ (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., $S_M = \{(\mu, \sigma)|\Delta P_M(\mu, \sigma) \geq \epsilon\}$ 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 $\pi_M$ is, by construction, within a KL range of $\delta$ from $\pi_0$, the agent can reach $S_M$ by updating its default policy to $\pi_M$ (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. Finally, Fig. 4 illustrates the granularity-complexity tradeoff: the granularity of the state representation, quantified as $-\log(\epsilon)$ (abscissa), controls the complexity capacity required to reach each state (ordinate). Finer-grained representations are generally more controllable. For a granularity level of $\epsilon = 0.1$ (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).

## 4 Discussion

We illustrated a novel approach to modelling 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 theoretical grounding for modelling how human, and non-human, agents select which features of their environment to attend to, and at what resolution, and which to ignore (Niv et al., 2015; Langdon et al., 2019). While several approaches for goal-directed state abstraction have been previously proposed (Li et al., 2006; Shah et al., 2021; Kaelbling, 1993; Zhang et al., 2020) our emphasis in this work is on principled theoretical perspective on the role of goals in shaping state representation learning by complexity constrained cognitive agents. Our quantification of policy complexity follows previous work applying information theoretic principles in reinforcement learning (Tishby & Polani, 2010; Rubin et al., 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; we accordingly posit that bounded agents should prefer, all else being equal, goals producing telic-controllable state representations, that balance control over the environment with with behavioral adaptability (cf. Klyubin et al. (2005)).

## References


David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L Littman, and Lawson LS Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3134–3142, 2019.

David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1639–1650. PMLR, 2020.

Nadav Amir, Reut Suliman-Lavie, Maayan Tal, Sagiv Shifman, Naftali Tishby, and Israel Nelken. Value-complexity tradeoff explains mouse navigational learning. *PLOS Computational Biology*, 16(12):e1008497, 2020.

Nadav Amir, Yael Niv, and Angela Langdon. 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*, 2023.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning*, pp. 373–382. PMLR, 2021.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*, 2022.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In *International conference on machine learning*, pp. 1515–1528. PMLR, 2018.

Leslie Pack Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pp. 1094–8. Citeseer, 1993.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pp. 128–135. IEEE, 2005.

Timothy A Krausz, Alison E Comrie, Ari E Kahn, Loren M Frank, Nathaniel D Daw, and Joshua D Berke. Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron*, 111(21):3465–3478, 2023.

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

Angela J Langdon, Mingyu Song, and Yael Niv. Uncovering the 'state': Tracing the hidden state representations that structure learning and decision-making. *Behavioural Processes*, 167:103891, 2019.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *AI&M*, 2006.

Gaia Molinaro and Anne G. E. Collins. A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 2023.

Paul S Muhle-Karbe, Hannah Sheahan, Giovanni Pezzulo, Hugo J Spiers, Samson Chien, Nicolas W Schuck, and Christopher Summerfield. Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, 111(23):3885–3899, 2023.

Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

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

Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in MDPs. *Decision Making with Imperfect Decision Makers*, pp. 57–74, 2012.

Dhruv Shah, Peng Xu, Yao Lu, Ted Xiao, Alexander Toshev, Sergey Levine, and Brian Ichter. Value function spaces: Skill-centric state abstractions for long-horizon reasoning. *arXiv preprint arXiv:2111.03189*, 2021.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle: Models, architectures, and hardware*, pp. 601–636. Springer, 2010.

Mianchu Wang, Yue Jin, and Giovanni Montana. Goal-conditioned offline reinforcement learning through state space partitioning. *Machine Learning*, pp. 1–31, 2024a.

Zizhao Wang, Caroline Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Building minimal and reusable causal state abstractions for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15778–15786, 2024b.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

## A Learning with telic states

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

$$\pi(a_i|o_1, a_1, ..., o_i). \tag{7}$$

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

$$e(o_i|o_1, a_1, ..., a_{i-1}). \tag{8}$$

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

$$P_\pi(o_1, a_1, ..., o_n, a_n) = P(o_1, a_1, ..., o_n, a_n|e, \pi) = \prod_{i=1}^{n} e(o_i|o_1, a_1, ..., a_{i-1})\pi(a_i|o_1, a_1, ..., o_i). \tag{9}$$

Typically, the environment is assumed to be fixed, and hence not explicitly parameterized in $P_\pi(h)$ above. As mentioned in the main text (section 2.3), the definition of telic states as goal-induced equivalence classes can now be extended to equivalence between policy-induced experience distributions as follows:

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \tag{10}$$

The question we are interested in can now be stated as follows: how can an agent learn an efficient policy for reaching a desired telic state? In other words, how can an agent increase the likelihood that its policy will generate experiences that belong 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 $\pi$:

$$\hat{P}_\pi(h) = \frac{|\{i : h_i = h\}|}{N}. \tag{11}$$

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

$$R = D_{KL}(P_i^\star || P_\pi) \tag{12}$$

where,

$$P_i^\star = arg \min_{P \in S_i} D_{KL}(P||P_\pi), \tag{13}$$

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

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta D_{KL}(P_i^\star || P_{\pi_\theta}), \tag{14}$$

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

# B Pseudocode for telic-controllable state representation learning algorithm

In this appendix we provide pseudocode for the telic-controllable state representation learning algorithm described in section 2.4 of the main text. The algorithm (1), makes use of an auxiliary procedure, FINDREACHABLESTATES (2), to find all reachable states, given the agent's goal, $g$, default policy, $\pi_0$, and policy complexity constraint, $\delta$. 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. It's main optimization step (line 3) can be implemented, e.g., using policy gradient over the information projection of $P_{\pi_0}$ on $S$.

**Algorithm 1** Telic-controllable state representation learning

**Input:** $\pi_0$: default policy, $g$: current goal, $\delta$: policy complexity capacity, $\epsilon$: sensitivity.
**Output:** $g'$: new goal such that $\mathcal{S}_{g'}$ is telic-controllable with respect to $\pi_0$ and $\delta$

$\mathcal{R} \leftarrow [P_{\pi_0}]_{\sim_g}$ ▷ initialize reachable state set
$g' \leftarrow g$ ▷ initialize new goal
**while** $\mathcal{R} \neq \mathcal{S}_{g'}$ **do**
  $\mathcal{R} \leftarrow$ FINDREACHABLESTATES$(\pi_0, g', \delta)$ ▷ 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 $P_{\pi_0}$ 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 $P_{\pi_0}$
    $S_M \leftarrow \{P : P \sim_g^{(\epsilon)} P_M\}$ ▷ $\epsilon$-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:** $\pi_0$: initial policy, $g$: goal, $\delta$: policy complexity constraint.
**Output:** all telic states in $\mathcal{S}_g$ reachable from $\pi_0$ by $\delta$-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$(\pi_0, g, \delta)$
  $\mathcal{R}_0 \leftarrow [P_{\pi_0}]_{\sim_g}$ ▷ 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**

# C Telic-complexity curves

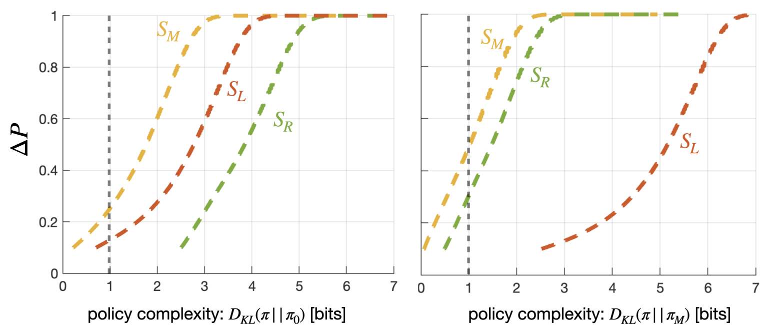


Figure 5: **Goal-complexity tradeoff curves:** the probability of reaching each telic state as a function of policy complexity. **Left:** an agent with a default policy $\pi_0 : (\mu_0, \sigma_0) = (0, 1)$ is unable to reach telic state $S_R$ with a complexity capacity limit of $\delta = 1$ (gray vertical 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.