States as goal-directed concepts: an epistemic approach to state-representation learning

Nadav Amir
nadav.amir@princeton.edu
Princeton University

Yael Niv
yael@princeton.edu
Princeton University

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

Abstract

Goals fundamentally shape how we experience the world. For example, when we are hungry, we tend to view objects in our environment according to whether or not they are edible (or tasty). Alternatively, when we are cold, we view the very same objects according to their ability to produce heat. Computational theories of learning in cognitive systems, such as reinforcement learning, use state-representations to describe how agents determine behaviorally-relevant features of their environment. However, these approaches typically assume ground-truth state representations that are known to the agent, and reward functions that need to be learned. Here we suggest an alternative approach in which state-representations are not assumed veridical, or even pre-defined, but rather emerge from the agent's goals through interaction with its environment. We illustrate this novel perspective using a rodent odor-guided choice task and discuss its potential role in developing a unified theory of experience based learning in natural and artificial agents.

1 Introduction

Concepts are the building blocks of mental representations, providing scaffolding for generalizations over individual objects or events. How do animals (including humans) form concepts and why do they form the particular ones that they do? These questions have a long history in both Eastern and Western philosophy (Hume, 1896; Wittgenstein, 1953; Rosch & Mervis, 1975; Siderits et al., 2011). More recently, computational cognitive science has started addressing a related question, namely, how do cognitive agents generate internal models of their environment, called "state-representations", generalizing over their experiences for efficient learning (Niv, 2019; Langdon et al., 2019; Song et al., 2022)? Here, we suggest that state-representations can be understood as conceptual frameworks formed by an agent in order to achieve particular goals (Dunne, 2004). Under this account, a concept, such as "fire", is formed because some set of interests, such as a desire for warmth, leads agents to construe particular entities as similar to each other in virtue of their efficacy in obtaining the goal of warming up. Utilizing this parallel between "concepts" and "states", we propose that state-representations should be understood in terms of the goals they subserve. To explore this hypothesis, we develop a formal framework for describing goal-dependent state-representations and illustrate its application by inferring animals' goals from empirically observed behavior in a well-studied odor-guided choice task.

2 Formal setting

2.1 Telic states as goal-equivalent experiences

We assume the setting of a perception-action cycle, i.e., streams 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 properties:

- 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:

$$\mathcal{S}_g = \Delta(\mathcal{H}) / \sim_g . \tag{1}$$

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.

2.2 Learning with telic states

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

$$\pi(a_i | o_1, a_1, ..., o_i). \tag{2}$$

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

$$e(o_i | o_1, a_1, ..., a_{i-1}). \tag{3}$$

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

$$P_\pi(o_1, a_1, ..., o_n, a_n) = P(o_1, a_1, ..., o_n, a_n | e, \pi) = \prod_{i=1}^{n} e(o_i | o_1, a_1, ..., a_{i-1}) \pi(a_i | o_1, a_1, ..., o_i). \tag{4}$$

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

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \tag{5}$$

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 the agent's policy be updated to increase its likelihood of generating experiences that belong to a certain desirable telic state,

$S_i \in \mathcal{S}_g$? To answer this question, we begin by writing down the empirical distribution of N experience sequences generated by policy π:

$$\hat{P}_\pi(h) = \frac{|\{k : h_k = h\}|}{N}. \tag{6}$$

We would like to estimate the probability that $\hat{P}_\pi(h)$ belongs to telic state S_i, and update π to increase this probability. A fundamental result from large deviation theory, known as Sanov's theorem (Cover, 1999), shows that this probability decays exponentially with a rate of

$$R = \min_{P \in S_i} D_{KL}(P||P_\pi). \tag{7}$$

Since R determines the probability that experiences sampled from P_π belong to telic state S_i, we refer to it as the *telic distance* from π to S_i. Assuming now the agent's policy can be expressed using some parameterization θ, the following policy gradient method updates π_θ in a way that minimizes the telic distance, i.e., maximizes the likelihood of generating experiences belonging to telic state 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 and,

$$P_i^\star = arg \min_{P \in S_i} D_{KL}(P||P_\pi), \tag{9}$$

is called *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. Equation 8 thus describes a general policy gradient method for learning with telic state representations.

2.3 Illustrative example: the two-armed bandit

To illustrate our proposed learning algorithm, we compute the goal-directed policy gradient for a fully-tractable bandit learning problem and show that, in this simple case, minimizing telic distance yields a commonly reported empirical choice strategy known as probability-matching. We consider a two-armed bandit in which the set of actions is defined as of choosing a left (L) or right (R) lever and the observations are winning (1) or losing (0):

$$\mathcal{A} = \{L, R\},\ \mathcal{O} = \{1, 0\}. \tag{10}$$

For simplicity we consider a past-independent policy, π_θ, that is parameterized by the probability of choosing action L:

$$\pi_\theta(L) = \theta,\ \pi_\theta(R) = 1 - \theta. \tag{11}$$

The environment e is specified by the probabilities of winning when choosing L or R, denoted p_L and p_R, respectively:

$$e(1|L) = p_L,\ e(0|L) = 1 - p_L;\ e(1|R) = p_R,\ e(0|R) = 1 - p_R. \tag{12}$$

The likelihood that an experience sequence, h, will be generated by the policy induced distribution P_{π_θ} can be expressed as:

$$P_{\pi_\theta}(h) = \theta^{N_L^h}(1-\theta)^{N_R^h} p_L^{N_{L,1}^h}(1-p_L)^{N_L^h - N_{L,1}^h} p_R^{N_{R,1}^h}(1-p_R)^{N_R^h - N_{R,1}^h}, \tag{13}$$

where N_L^h, N_R^h are the number of times the agents selected the L and R actions, respectively, and $N_{L,1}^h, N_{R,1}^h$ are the number of "win" observations following L and R choices, respectively. For simplicity, we assume that the agents goal is to reach a specific number of wins, so that two policies are equivalent if and only if the expected number of wins obtained by following both is equal:

$$\pi_{\theta_1} \sim_g \pi_{\theta_2} \iff \mathbb{E}_{P_{\pi_{\theta_1}}(h)}\left(\sum_i^N \mathbf{1}_{h_i=1}\right) = \mathbb{E}_{P_{\pi_{\theta_2}}(h)}\left(\sum_i^N \mathbf{1}_{h_i=1}\right), \tag{14}$$

where $N = N_L^h + N_R^h$ is the total number of action/observation pairs and $\mathbf{1}_{h_i=1}$ denotes an indicator function which is one if the ith observation in h is 1 and zero otherwise. Thus, for every $j = 1, ..., N$, the telic state S_j is defined simply as the set of all experience distributions with an expected number of wins equal to j:

$$S_j = \{P(h) : \text{ s.t. } \mathbb{E}_{h\sim P}\left(\sum_i^N \mathbf{1}_{h_i=1}\right) = j\}. \tag{15}$$

The telic distance (Eq.7) between a policy π_θ and S_j is given by:

$$D_{KL}(S_j||P_{\pi_\theta}) = \min_{P\in S_j} D_{KL}(P||P_{\pi_\theta}) = \sum_h P_j^\star(h) \log \frac{P_j^\star(h)}{P_{\pi_\theta}(h)}, \tag{16}$$

where $P_j^\star(h)$ is the distribution in S_j closest to $P_\pi(h)$ in the KL sense, as defined in Eq. 9 above. Using Eq. 13 we can compute the telic distance gradient:

$$\nabla_\theta D_{KL}(S_i||P_{\pi_\theta}) = \frac{\sum_h P_j^\star(h) N_R^h}{1-\theta} - \frac{\sum_h P_j^\star(h) N_L^h}{\theta}, \tag{17}$$

so that the optimal policy for reaching j wins, $\pi_{\theta_j^*}$, is given by:

$$\theta_j^* = \mathbb{E}_{h\sim P_j^\star}\left(\frac{N_L^h}{N_L^h + N_R^h}\right). \tag{18}$$

In words, the policy maximizing the likelihood of reaching telic state S_j, is one matching the expected choice probability of $P_j^\star(h)$. Interestingly, a similar "probability-matching" strategy was found in human iterated binary choice behavior (Erev & Barron, 2005).

2.4 Transition sensitive goals and the flow of experience

So far, we considered goals that are sensitive to individual action or observation counts within experience sequences. Indeed, the standard reinforcement learning framework can be viewed as special cases of ours, under a goal of maximizing a (possibly discounted) sum of individual observations deemed as rewarding by the agent or task-designer. Real cognitive agents however, typically pursue more ecological goals, that reflect complex preferences over higher-order statistics of experience, beyond the accumulation of local reward signals. For example, people engaged in activities such as games, artistic creation, sports, meditation etc., often describe the goal of such pursuits as entering a state of "flow"(Csikszentmihalyi & Csikszentmihalyi, 1992), described in terms of matching skill and challenge levels. While such activities may not typically be thought of as goal directed, and thus do not readily lend themselves to standard reward-driven learning modelling frameworks, they can be expressed in the current framework in terms of higher order experience correlations structures. For example, consider an agent that prefers certain observations (or actions), but only when they follow certain actions (or observations). This "second-order" preference structure can be described by a class of goals in which experiences are ordered according to the weighted sum of possible observation-action and action-observation transition counts:

$$g_{\alpha,\beta}(h) = \sum_{i=1}^{|\mathcal{O}|}\sum_{j=1}^{|\mathcal{A}|}(N_{ij}^h \alpha_{ij} + M_{ji}^h \beta_{ji}), \tag{19}$$

where N_{ij}^h denotes the number of transitions between observation $o_t = i$ and action $a_{t+1} = j$, and M_{ji}^h the number of transitions between action $a_{t-1} = j$ and observation $o_t = i$, in a given experience sequence $h = o_1, a_1, ..., o_n, a_n$. Under this goal, experience sequences sharing the same empirical transition frequencies will be deemed equivalent by the agent, with the α_{ij} and β_{ji} parameters determining the relative weight of different action-observation or observation-action transition, respectively. As a concrete example, consider a setting where observations depend only on the immediately preceding action and actions depend only on the immediately preceding observation. The

environment can thus be expressed as $e(o_i|o_1, a_1, ..., o_{i-1}, a_{i-1}) = e(o_i|a_{i-1})$, and the policy of the agent as $\pi(a_i|o_1, a_1, ..., o_i) = \pi(a_i|o_i)$. The experience distribution induced by the environment e and the policy π can be factorized in this case as:

$$P_\pi(h) = e(o_1)\pi(a_1|o_1)\prod_{i=2}^{n} e(o_i|a_{i-1})\pi(a_i|o_i). \tag{20}$$

This distribution can be parameterized using the switching probabilities $e_{ji} \equiv e(o_t = i|a_{t-1} = j)$, with an initial observation distribution, $e_i = e(o_1 = i)$, and $\pi_{ij} \equiv \pi(a_t = j|o_t = i)$ for $i = 1, ..., |\mathcal{O}|$ and $j = 1, ..., |\mathcal{A}|$. Using this notation, we can express the log-probability of experience h given policy π (and environment e) as:

$$\log P_\pi(h) = \log\left(e(o_1)\prod_{i=1}^{|\mathcal{O}|}\prod_{j=1}^{|\mathcal{A}|}\pi_{ij}^{N_{ij}^h} e_{ji}^{M_{ji}^h}\right) = \log e(o_1) + \sum_{i=1}^{|\mathcal{O}|}\sum_{j=1}^{|\mathcal{A}|}(N_{ij}^h \log(\pi_{ij}) + M_{ji}^h \log(e_{ji})). \tag{21}$$

Assuming sufficiently long sequences, $N_{ij}^h \approx \frac{n}{2}\pi_{ij}$ and $M_{ji}^h \approx \frac{n}{2}e_{ji}$, the goal $g_{\alpha,\beta}(h)$ can be approximated by:

$$g_{\alpha,\beta}(h) \approx \frac{n}{2}\sum_{i=1}^{|\mathcal{O}|}\sum_{j=1}^{|\mathcal{A}|}(\pi_{ij}\alpha_{ij} + e_{ji}\beta_{ji}) = -\frac{n}{2}(H(\pi, \pi_\alpha) + H(e, e_\beta)) \tag{22}$$

where $H(p, q) = -\mathbb{E}_p[\log q]$ denotes the cross entropy of distribution q relative to p, and the distributions π_α and e_β are defined as:

$$\pi_\alpha(a_t = j|o_{t-1} = i) = \exp(\alpha_{ij}), \tag{23}$$

and

$$e_\beta(o_t = i|a_{t-1} = j) = \exp(\beta_{ji}), \tag{24}$$

with the following normalization constraints on α_{ij} and β_{ji}:

$$\sum_{j=1}^{|\mathcal{A}|}\exp(\alpha_{kj}) = \sum_{i=1}^{|\mathcal{O}|}\exp(\beta_{li}) = 1, \text{ for } k = 1, ..., |\mathcal{O}| \text{ and } l = 1, ..., |\mathcal{A}|. \tag{25}$$

Thus, for a fixed environment, a goal of maximizing a weighted sum of specific action-observation and observation-action transitions, $g_{\alpha,\beta}(h)$, corresponds to minimizing the cross entropy between the true policy π, and a reference one π_α, induced only by the observation-action transition weight parameters, α_{ij}. Such a transition-sensitive goal therefore drives agents to optimize an intrinsically emergent policy complexity cost (Amir et al., 2020; Lai & Gershman, 2024; Arumugam et al., 2024). Extending this analysis to preferences over longer flows of experience, i.e., action-observation subsequences, can provide a formal approach to learning in cognitive systems, driven by ecological goals that are sensitive to intricate statistical features of experience, beyond accumulation of local events designated as "rewarding" by an external task or agent designer.

2.5 Closing the loop: telic state conditioned policies

Above we have assumed that the policy depends on the full past experience but this assumption can be relaxed. Within the current framework, we assume that the agent maintains an estimate of the most likely telic state it is currently in and updates it at each time point. Concretely, given a goal g and a past experience sequence at time t, $h_t = o_1, a_1, ..., o_t$, the agent can estimate its current telic state, i.e., the equivalence class of the experience distribution most likely to have generated h_t:

$$\hat{S}_t(h_t) = [\arg\max_{P \in \Delta(\mathcal{H}_t)} P(h_t)]_{\sim_g}. \tag{26}$$

The policy can now be expressed in terms of the estimated telic state:

$$\pi(a_t|\hat{S}_t(h)), \tag{27}$$

so that in choosing actions the agent generalizes over past experiences that are estimated to originate from the same telic state. Since the borders between telic states are determined by the goal, the same experience may be assigned to different telic states under different goals. This clustering of past experience into estimated telic states is lossy: it ignores goal-irrelevant information, and is not necessarily Markovian. Thus, while it may not be optimal in the Bayesian sense, it provides a self contained account of how goals, i.e., preferences over experience distributions, generate intrinsic (telic) state representation, which in turn provide a foundation for action selection and learning.

3 Results: goal inference in an odor-guided choice task

To illustrate the application of our proposed framework to behavioral data analysis, we infer the preference profile, i.e., goal, of individual animals from their empirical choice behavior and use it to explain reaction time in a well-studied perceptual decision making task (Roesch et al., 2006). Briefly, rats were trained to sample an odor at a central odor port, before responding by nose-poking in one of two fluid wells. The odor stimulus provided a cue for which of two wells would be associated with delivery of a certain amount of sucrose liquid. Two of the odors signalled "forced choice" trials, one indicating that the liquid will be available in the left well, and one indicating the right well. A third odor—"free choice"—indicated liquid availability in either well. After liquid delivery, or choice of a non-indicated well, the rat waited for a cue indicating the start of the next trial. Importantly, if a "valid" well was chosen on any trial (i.e., the indicated well on forced-choice trials, or either well on free-choice trials), the delay to and amount of liquid was determined by the side of the well, not the odor. Unsignaled to the animal, in each block of the task, one well delivered either at a shorter delay or a larger amount than the other well; amount and delay contingencies changed between blocks during a session. We denote the set of possible observations and actions in the task as follows:

$$\begin{aligned}\mathcal{O} = \{&\textit{Left Odor, Right Odor, Free Odor,}\\ &\textit{Long Delay, Short Delay, Small Amount, Big Amount}\},\\ \mathcal{A} = \{&\textit{Right Poke, Left Poke, Wait for Cue}\}.\end{aligned}$$

We define an experience for this task as a sequence of trials, each consisting of an observation-action pair, for example:

$$h = LO, LP, LD, WC, FO, RP, SD, WC, ... \tag{28}$$

where elements of $\mathcal{O}$ and $\mathcal{A}$ are denoted by their initials.

While goals are generally defined as preferences over experience *distributions*, here we shall consider a simplified special case in which the goal is defined in terms of order over individual experiences. This is indeed a special case, since any individual experience can be thought of as a delta function distribution concentrated around itself. Specifically, we define a parameterized family of goals that assign a score to an experience h based on the weighted sum, normalized by length, of the difference between the number of big vs. small amount observations, short vs. long delay observations, and right vs. left nose poke actions appearing in it:

$$g_\beta(h) = \frac{1}{N^h}\big(\beta_1(N^h_{BA} - N^h_{SA}) + \beta_2(N^h_{SD} - N^h_{LD}) + \beta_3(N^h_{RP} - N^h_{LP})\big), \tag{29}$$

where $N^h_{BA}, N^h_{SA}, N^h_{SD}$ and N^h_{LD} denote the number of of *Big Amount, Small Amount, Short Delay* and *Long Delay* observations, respectively, and N^h_{RP} and N^h_{LP} denote the number of *Right Poke* and *Left Poke* actions, respectively, in experience h. The normalization by the total experience length, N^h, renders the goal score scale-invariant, allowing experiences of different lengths to be compared in terms of their alignment with goal g_β (see Eq. 31 below). The parameters: β_1, β_2, and β_3, determine the relative weight of the corresponding difference between action or observation counts in determining each animal's goal. Thus, we assume that the preference profile, i.e., goal, of each rat is characterized by its unique 3D parameter vector $\beta = (\beta_1, \beta_2, \beta_3)$. While our formalism is general with respect to the form of the goal, this particular goal parameterisation allows us to

monitor preference for earlier rewards, larger rewards, and side biases – known characteristics of animal behavior in this task.

To illustrate our framework in the context of this task (Fig. 2), we define a *telic state-trajectory* as a sequence of estimated telic states, $\{\hat{S}_t^\beta(h)\}_{t=1}^n$, where $\hat{S}_t^\beta(h)$ consists of all experiences of length t that are g_β-equivalent to the first t trials of the experience h:

$$\hat{S}_t^\beta(h) = \{h' \in \mathcal{H}_t : g_\beta(h') = g_\beta(h_{1:t})\}, \tag{30}$$

where $h_{1:t} = o_1, a_1, ..., o_t, a_t$ denotes sub-sequence consisting of the first t observations and actions in h. We quantify the alignment of an empirical experience h with a goal g_β using the following Goal Alignment Coefficient (GAC), defined simply as the g_β score of h:

$$GAC(h; g_\beta) = g_\beta(h). \tag{31}$$

For a given experience, h, the GAC measures the average increase in the goal value (Eq. 29) per trial in h. We used the GAC, to estimate the parameters β^* that maximize the alignment between a given empirical set of M experience sequences $\{h_j\}_{j=1}^M$ and the goal g_{β^*}, subject to a regularization constraint on β:

$$\beta^* = \arg\max_\beta \sum_{j=1}^{M} GAC(h_j; g_\beta) \text{ s.t } \|\beta\|^2 = 1, \tag{32}$$

where we impose the regularization constraint $\|\beta\|^2 = \sum_{i=1}^3 \beta_i^2 = 1$ on the weight parameters to fix the scale of g_β. In other words, given the empirical experience sequences of an individual animal, and a class of β-parameterized goals, we estimated the parameter values, β^* such that the animal's behavior is maximally aligned with the goal g_{β^*}. The β^* values for all animals are plotted in Fig. 1 (orange histograms). As a baseline for comparison, we also plotted β^* values for simulated animals using the same odor observations and liquid-outcome contingencies as the empirical data but with randomly selected left or right action choices (blue histograms). This analysis revealed that goals fit to the empirical behavior had significantly larger weights on the difference between the number of short vs. long delays and big vs. small liquid amounts compared to goals fitted to simulated random behavior. Fig. 2 shows the telic state trajectories for empirical and simulated experience sequences, and illustrates how this novel analysis may be used to explain features of behavioral learning. We plotted the optimized goal scores, $g_{\beta^*}(h)$, for the first 1000 trials of individual animals using empirical and random choice simulated data (Fig. 2A, orange and blue lines respectively). Optimized goal alignment coefficients were significantly higher for empirical compared to simulated animals (Fig. 2B, orange and blue bars, respectively), demonstrating the sensitivity of the GAC as as general-purpose measure of goal-directed behavior. Finally, (Fig. 2C) shows choice reaction time as a function of telic state and trial number for an individual animal, whose state trajectory is marked by the gradient-colored line in panel (A), with darker colors corresponding to earlier trials. Reaction times were significantly correlated with telic state (top) despite not being consistently related to trial number (bottom). Overall, this analysis illustrates the potential use of telic state-based analysis to uncover new features of behavioral learning.

4 Discussion

The need for a formal theory of learning in cognitive systems that centers on agent's goals has recently been called into attention (Molinaro & Collins, 2023). The current work provides a step in this direction, leveraging a novel definition of telic states as equivalence classes of experiences distributions with respect to goals. While the notion of states as equivalence classes is not new (Minsky, 1967), we suggest that goals, defined as preference relations over experience distributions (Bowling et al., 2022), provide a natural foundation for state representation learning in cognitive systems. Our approach is related to recent results in preference-based reinforcement learning (Wirth et al., 2017; Carr et al., 2024) and goal directed state-abstraction (Li et al., 2006; Abel et al., 2019). However, unlike previous accounts that assume a pre-defined "ground" state representation or reward

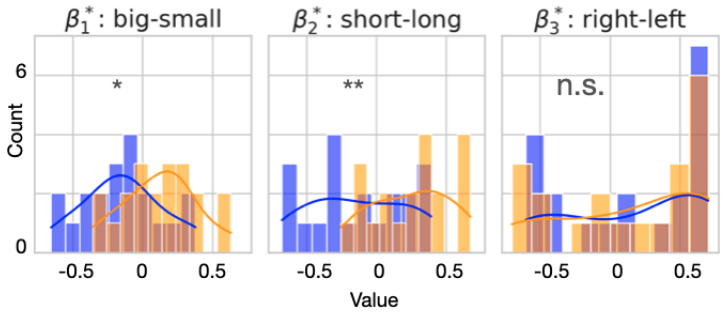

Figure 1: **Optimized weight parameters for liquid amount (left), delay duration (center) and side choice (right) preferences.** Optimized β values maximizing the goal-alignment coefficient of empirical experience sequences (orange) are significantly larger than those of simulated random actions yoked to the observation experience sequences of each animal (blue) for big vs. small amount and short vs. long delay but not for right vs. left nose pokes. Solid lines show Gaussian kernel density distribution estimates. Asterisks indicate significance levels (paired t-test, $^*p < 0.01$; $^{**}p < 0.001$).

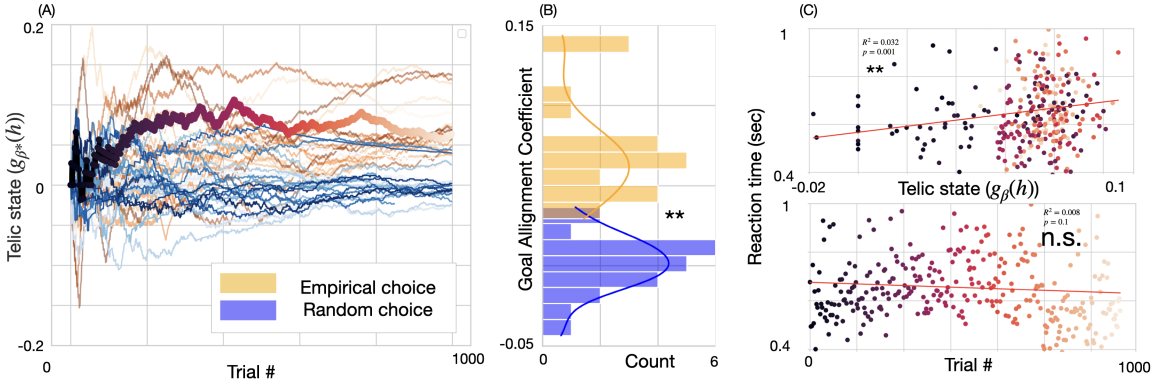

Figure 2: **Telic state trajectories and choice reaction times. (A)** Telic state trajectories for the first 1000 trials, defined as the optimized goal function score $g_{\beta*}(h)$ (ordinate) along consecutive experience sub-sequences (abscissa). Each trajectory represents a sequence of trials experienced by an individual rodent, with orange ones corresponding to empirical behavior and blue ones to simulated behavior using the same observation sequences experiences by the real animals but with randomly chosen actions. **(B)** The corresponding Goal Alignment Coefficient (GAC) histograms show that real animals (orange bars) achieve significantly higher telic states than simulated ones (blue bars) (paired t-test, $p < 10^{-4}$) (B). **(C)** Reaction times as a function of telic state (top) and trial number (bottom) for an individual rat, indicated by the gradient colored line in panel (A). Each dot represents a single choice trial, with lighter colors indicating later serial trial positions. For this animal, higher telic states were correlated with longer reaction times overall (top), despite the lack of significant correlation between trial number and reaction time (bottom).

function (or both), we posit that state representations are inherently goal directed, obviating the need to assume given states and reward functions. While reward functions can be extended to experience-based learning settings (Lu et al., 2023), we suggest that replacing rewards by goals has several advantages. Beyond its conceptual parsimony, a goal-based framework may be used to bridge natural and artificial perspectives on state representation learning by using telic states to describe both the evaluative and the descriptive aspects of learning models. Thus, whereas previous goal-conditioned reinforcement learning frameworks typically treat goals as a special subset of states (Kaelbling, 1993) or as a parameterization of the reward function within a given state space (Schaul et al., 2015), our framework views goals and states as distinct yet coupled constructs. This coupling

between goals and telic state representations suggests a "double dissosoacition" between experiences and telic states: on the one hand, the same telic state can be obtained via two different, yet equally desirable, experiences; while on the other hand, the same experience can result in different telic states under different goals, as goals determine which aspects of experience are important in a given context. Thus, unlike standard Bayesian or decision theoretic models that attempt to explain behavioral response variability for identical stimuli as noise or suboptimality, our framework can account for such variability as epistemic optimally under switching goals, e.g., engagement and relaxation (Ashwood et al., 2022). From a neurocognitive perspective, our framework predicts that state representations in the brain, such as those believed to be encoded by neuronal activity patterns in the hippocampus (Crivelli-Decker et al., 2023) and prefrontal areas such as the orbitofrontal cortex (Schuck et al., 2018), should be sensitive to agents' goals and reflect their "position" within a corresponding telic state space (De Martino & Cortese, 2023).

In a sense, our approach shifts the burden from explaining the origin of state representations to explaining the origin sensory-motor ones, as it assumes given sets of possible actions and observations. However, sensory-motor affordances can also be explained in terms of underlying goals, as evinced by recent studies on motivated perception (Balcetis & Dunning, 2006; Leong et al., 2019) and task-conditioned action representation (Fogassi et al., 2005; Aberbach-Goodman et al., 2022; Ahn et al., 2022). Furthermore, as the biological nervous systems has arguably been evolutionary shaped to detect and respond to goal relevant information (Lettvin et al., 1959), a hierarchy of goals, operating at different timescales, may generate a hierarchy of state representations at multiple levels of temporal abstraction (Sutton et al., 1999; Shah et al., 2021).

Finally, our framework can provide a formal setting for addressing the fundamental problem of goal selection, namely, where do goals come from and how do agents choose between different goals the first place? By coupling goals and state representations, the current framework suggests that goals may be formed based on properties of the state representations they induce. For example, all else being equal, agents may prefer state representations that are more "controllable", i.e., ones that make it easier to explore all regions within their space of telic states, and adjust their goals accordingly (Amir et al., 2024; Klyubin et al., 2005). Ultimately, our framework provides a step towards a formal account of how goals can structure experience in cognitive systems.

Acknowledgements

This work was supported by grant U01DA050647 from the National Institute on Drug Abuse and the Intramural Research Program of the National Institute of Mental Health under ZIAMH002983 (to AJL).

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.

Shahar Aberbach-Goodman, Batel Buaron, Liad Mudrik, and Roy Mukamel. Same action, different meaning: neural substrates of action semantic meaning. *Cerebral Cortex*, 32(19):4293–4303, 2022.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

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, Stas Tiomkin, and Angela Langdon. Learning telic-controllable state representations. *arXiv preprint arXiv:2406.14476*, 2024.

Dilip Arumugam, Mark K Ho, Noah D Goodman, and Benjamin Van Roy. Bayesian reinforcement learning with limited cognitive load. *Open Mind*, 8:395–438, 2024.

Zoe C Ashwood, Nicholas A Roy, Iris R Stone, International Brain Laboratory, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, 2022.

Emily Balcetis and David Dunning. See what you want to see: motivational influences on visual perception. *Journal of personality and social psychology*, 91(4):612, 2006.

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

Jonathan Colaço Carr, Prakash Panangaden, and Doina Precup. Conditions on preference relations that guarantee the existence of optimal policies. In *International Conference on Artificial Intelligence and Statistics*, pp. 3916–3924. PMLR, 2024.

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

Jordan Crivelli-Decker, Alex Clarke, Seongmin A Park, Derek J Huffman, Erie D Boorman, and Charan Ranganath. Goal-oriented representations in the human hippocampus during planning and navigation. *Nature communications*, 14(1):2946, 2023.

Mihaly Csikszentmihalyi and Isabella Selega Csikszentmihalyi. *Optimal experience: Psychological studies of flow in consciousness*. Cambridge university press, 1992.

Benedetto De Martino and Aurelio Cortese. Goals, usefulness and abstraction in value-based choice. *Trends in Cognitive Sciences*, 27(1):65–80, 2023.

John D Dunne. *Foundations of Dharmakirti's philosophy*. Simon and Schuster, 2004.

Ido Erev and Greg Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912, 2005.

Leonardo Fogassi, Pier Francesco Ferrari, Benno Gesierich, Stefano Rozzi, Fabian Chersi, and Giacomo Rizzolatti. Parietal lobe: from action organization to intention understanding. *Science*, 308 (5722):662–667, 2005.

David Hume. *A treatise of human nature*. Clarendon Press, 1896.

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. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753. Springer, 2005.

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.

Yuan Chang Leong, Brent L Hughes, Yiyu Wang, and Jamil Zaki. Neurocomputational mechanisms underlying motivated seeing. *Nature human behaviour*, 3(9):962–973, 2019.

Jerome Y Lettvin, Humberto R Maturana, Warren S McCulloch, and Walter H Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959.

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

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, Zheng Wen, et al. Reinforcement learning, bit by bit. *Foundations and Trends® in Machine Learning*, 16(6): 733–865, 2023.

Marvin Minsky. Computation: Finite and infinite machines prentice hall. *Inc., Engelwood Cliffs, NJ*, 1967.

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

Yael Niv. Learning task-state representations. *Nature Neuroscience*, 22(10):1544–1553, 2019.

Matthew R Roesch, Adam R Taylor, and Geoffrey Schoenbaum. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron*, 51(4):509–520, 2006.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International conference on machine learning*, pp. 1312–1320. PMLR, 2015.

Nicolas W Schuck, Robert Wilson, and Yael Niv. A state representation for reinforcement learning and decision-making in the orbitofrontal cortex. In *Goal-directed decision making*, pp. 259–278. Elsevier, 2018.

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.

Mark Siderits, Tom JF Tillemans, and Arindam Chakrabarti. *Apoha: Buddhist nominalism and human cognition*. Columbia University Press, 2011.

Mingyu Song, Yuji K Takahashi, Amanda C Burton, Matthew R Roesch, Geoffrey Schoenbaum, Yael Niv, and Angela J Langdon. Minimal cross-trial generalization in learning the representation of an odor-guided choice task. *PLoS Computational Biology*, 18(3):e1009897, 2022.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017.

Ludwig Wittgenstein. *Philosophical investigations*. Wiley-Blackwell, New York, NY, USA, 1953.