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

Nadav Amir[1*] **Yael Niv**[1] **Angela Langdon**[2]
[1]Princeton University [2]National Institute of Mental Health
{nadav.amir,yael}@princetn.edu
angela.langdon@nih.gov

Abstract

Our 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 may view the very same objects according to their ability to produce heat. Computational theories of learning in cognitive systems, such as reinforcement learning, use the notion of "state-representation" to describe how agents decide which features of their environment are behaviorally-relevant and which can be ignored. However, these approaches typically assume "ground-truth" state representations that are known by 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 by inferring the goals driving rat behavior in an odor-guided choice task and discuss its implications for developing, from first principles, an information-theoretic account of goal-directed state representation learning and behavior.

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 [10, 23, 19, 20]. More recently, computational learning theories have 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 [17, 12]? Here, we suggest that state-representations can be understood as concepts, formed by an agent in order to achieve particular goals [7]. 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. Drawing 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

We assume the setting of an observation-action cycle, i.e., an agent receiving an observation from its environment and subsequently performing an action, then receiving a new observation and so on. 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: $h = o_1, a_1, o_2, a_2, ..., o_n, a_n$.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

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. [4], 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-induced 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 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 with respect to their goal. Furthermore, since different states are, by definition, non-equivalent with respect to $\succeq_g$, the goal also determines whether a transition between any two states brings the agent in closer alignment to, or further away from its goal.

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

We illustrate our proposed framework by inferring goal alignment from empirical behavior in a well-studied odor-guided choice task [18]. 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:

$$h = LO, LP, LD, WC, FO, RP, SD, WC, ...$$

where elements of $\mathcal{O}$ and $\mathcal{A}$ are denoted by their initials. For simplicity, we consider a simple parameterized family of goals, defined as the weighted sum of the differences between the number of right vs. left nose pokes, short vs. long delays, and big vs. small amounts, appearing in an individual experience:

$$g_\beta(h) = \beta_1(N^h_{BA} - N^h_{SA}) + \beta_2(N^h_{SD} - N^h_{LD}) + \beta_3(N^h_{RP} - N^h_{LP}), \tag{2}$$

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 parameters $\beta_1\beta_2$, and β_3, determine the relative weight of the corresponding difference between action or observation counts in determining the animal's goal. While our formalism is general with respect to the form of the goal, this particular goal family allows us to monitor preference for earlier rewards, larger rewards, and side biases – all characteristics of animal behavior in this task. For an empirical experience $h = o_1, a_1, ..., o_n, a_n$ consisting of n trials, we define the corresponding *state-trajectory* as the sequence of states $\{S_t^\beta\}_{t=1}^n$ where S_t^β consists of all t-trial histories that are g_β-equivalent to the first t trials of h:

$$S_t^\beta = \{h' \in \mathcal{H}_t : g_\beta(h') = g_\beta(o_1, a_1, ..., o_t, a_t)\}. \tag{3}$$

Given a goal, g_β, and an empirical experience, h, we quantify the alignment between h and g_β using the following Goal Alignment Coefficient (GAC):

$$GAC_\beta(h) = g_\beta(h)/n. \tag{4}$$

For a given experience, h, the GAC measures the average increase in the goal value (Eq.2) for each trial in h. We used the GAC, to estimate the parameters β^* that maximize the alignment between a given empirical set of histories $\{h_j\}_{j=1}^N$ and the goal g_{β^*}, subject to a regularization constraint on β:

$$\beta^* = \arg\max_\beta \sum_{j=1}^N GAC_\beta(h_j) \text{ s.t } \|\beta\|^2 = 1, \tag{5}$$

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 histories 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. Next, we plotted the value of g_β for the state trajectories defined by cumulative trial sequences of the empirical and simulated histories (Fig.2 top, orange and blue lines respectively). We compared the trajectories obtained using the optimized weights, β^*, (left) with those obtained using random weights sampled from a unit normal distribution, $\beta \sim \mathcal{N}(0, 1)$, (right). We also plotted the GACs for the empirical and simulated animals (Fig.2 bottom, orange and blue histograms), comparing optimized (left) and random (right) weights. Weight optimization significantly increased the goal-alignment for the empirical but not for the simulated animals, demonstrating the sensitivity of the GAC as as general-purpose measure of goal-directed behavior.

4 Discussion

4.1 Relation to previous work

The need for a formal theory of learning in cognitive systems that centers on agent's goals has recently been called into attention [16]. The current work provides a step in this direction, based on the definition of states as equivalence classes over experience sequences with respect to goals. While the notion of states as equivalence classes is not new [15], we suggest that goals, defined as preference relations over experience sequences [4], provide an natural epistemic foundation for constructing state representations in cognitive systems by generalizing over goal-equivalent experiences. Our approach differs from previous theoretical accounts of goal-directed state abstraction in important ways. First, while previous accounts typically assume a pre-given "ground" state representation and reward function and describe how various generalizations across these facilitate planning and decision-making in certain contexts (see [14] for a review), our approach does not assume any a-priori state representation or reward structure but rather posits that state representations are fully goal dependent. Thus, whereas previous goal-conditioned reinforcement-learning models view goals as a subset of future states that the agent wants to reach [11], our framework goes in the opposite direction by starting with the goal and deriving the state-representation that best alligns with it. In some sense, our approach "shifts the burden" from explaining state representations to explaining sensory-motor ones,

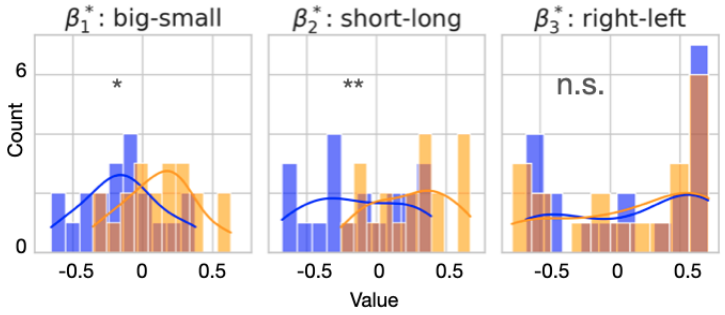

Figure 1: **Optimized weight parameters for liquid amount (left), delay duration (center) and side choice (right) preferences.** Optimized β values maximizing the goal-alignment coefficient of empirical histories (orange) are significantly larger than those of simulated random actions yoked to the observation histories of each animal (purple) 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, $^*p < 0.01$; $^{**}p < 0.001$).

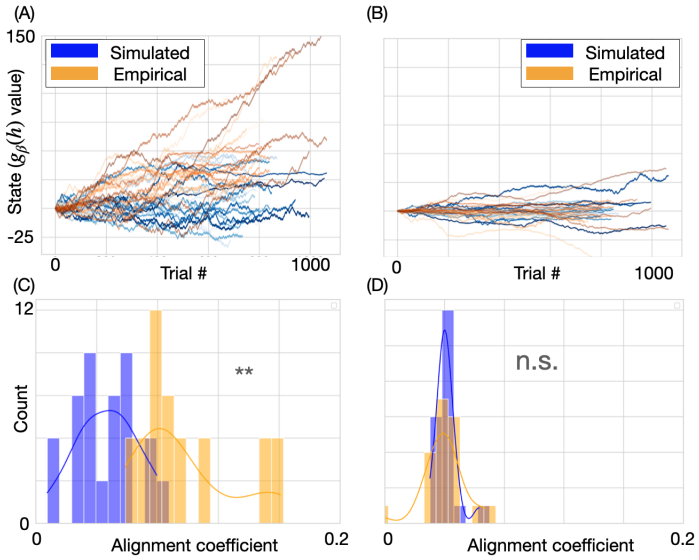

Figure 2: **State trajectories (top) and Goal Alignment Coefficients (bottom) for optimized weights (left) and random weights (right).** Top: State trajectories for optimized (A) and random (B) β weights show the state, defined as the cumulative weighted sum g_β for individual animals (ordinate) for different lengths of trial histories (abscissa). State trajectories for real animals (orange) reach larger $g_\beta(h)$ values than those of random simulated animals (blue). However, this is only true when using β^* (A), not with random values of β (B).Bottom: histograms of corresponding Goal Alignment Coefficients (GACs) of all animals (real in orange, simulated in blue) for optimized (left) and random (right) β weights. Optimized GACs for empirical trajectories were significantly larger than for simulated ones (paired t-test, $p < 10^{-4}$), but this is true only for optimized weights (C), not for random weights (D).

as it assumes a pre-defined set of possible actions and observations available to the agent. However, we argue that sensory-motor representations themselves may be similarly explained in terms of underlying goals, as evinced by recent studies on “motivated perception" [3, 13] and goal-conditioned action representations [9, 1]. Another potential challenge to our framework is that the definition of goal-states as equivalence classes appears inconsistent with reports that human preference in economic choices sometimes exhibit intransitively [22]. However, since decision making relies on sampling from experience distributions, choice anomalies such as preference intransitiviy can be explained in the current framework in terms of random sample deviations (see Supplementary material for details). Ultimately, our framework provides a step towards a formal, epistemological account of how, and why animals represent their environment in particular ways, and not others [21].

5 Supplementary material

5.1 Learning with goal directed states

So far, we have only outlined a framework for deriving goal-directed state representations and illustrated how it can be used to infer goals which best explain the observed behavior. It is natural to ask how can an agent use these goal-derived representations to guide goal-directed behavior. For this, we define a *policy*, π as a distribution over actions given the past experience sequence and current observation:

$$\pi(a_i|o_1, a_1, ..., o_i). \tag{6}$$

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

$$e(o_i|o_1, a_1, ..., a_{i-1}). \tag{7}$$

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

$$P_\pi(o_1, a_1, ..., o_n, a_n) = P(o_1, a_1, ..., o_n, a_n|e, \pi) = \prod_{i=1}^{n} e(o_i|o_1, a_1, ..., a_{i-1})\pi(a_i|o_1, a_1, ..., o_i). \tag{8}$$

Our definition of states as goal-induced equivalence classes can now be extended to equivalence between policy-induced experience distributions as follows:

$$\pi_1 \sim_g \pi_2 \iff P_{\pi_1} \sim_g P_{\pi_2}. \tag{9}$$

Typically, the environment is assumed to be fixed, and hence not explicitly parameterized in $P_\pi(h)$ above, and the question we are interested in is: how can an agent learn an efficient policy for achieving its goal? In the current framework, we interpret this question as how can an agent increase the likelihood that its policy will generate experiences that belong to a certain goal state S_i. In other words, assuming the agent follows a policy π, and experiences a sample of N sequences, $h_1, ..., h_N$, we would like to know how can the agent modify its policy such that the empirical distribution of these samples belongs to a certain goal-state S_i. Denoting the empirical distribution of sequences by $\hat{P}_\pi$, i.e., $\hat{P}_\pi(h) = \frac{|\{i:h_i=h\}|}{N}$, the answer to this question is given, in the asymptotic limit, by the following result from the theory of large-deviations known as Sanov's theorem:

$$\lim_{N\to\infty} \frac{1}{N} \log P_\pi^{(N)}(\hat{P}_\pi \in S_i) = -D_{KL}(P^\star||P_\pi), \tag{10}$$

where,

$$P_i^\star \equiv arg \min_{P\in S_i} D_{KL}(P||P_\pi), \tag{11}$$

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 [5]. In other words, the probability that experience sequences sampled from P_π will belong to a particular goal-state S_i is determined by the KL-divergence between the information projection of P_π onto S_i and P_π itself. Assuming now the agent follows a policy π_θ, parameterized by some θ, the following goal-directed gradient method can be used to update the policy π_θ to one that is more likely to generate experience sequences leading the agent to the goal-state S_i:

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta D_{KL}(P_i^\star||P_{\pi_\theta}). \tag{12}$$

5.1.1 Illustrative example - probability matching in 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 goal-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):

$$\mathcal{A} = \{L, R\},\ \mathcal{O} = \{1, 0\}. \tag{13}$$

The policy, π_θ, is parameterized by the probability of choosing action L:

$$\pi_\theta(L) = \theta,\ \pi_\theta(R) = 1 - \theta. \tag{14}$$

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

$$e(1|L) = p_L,\ e(0|L) = 1 - p_L;\ e(1|R) = p_R,\ e(0|R) = 1 - p_R. \tag{15}$$

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

$$P_{\pi_\theta}(h) = \theta^{N_L^h}(1-\theta)^{N_R^h} p_L^{N_{L,1}^h}(1-p_L)^{N_L^h - N_{L,1}^h} p_R^{N_{R,1}^h}(1-p_R)^{N_R^h - N_{R,1}^h}, \tag{16}$$

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 maximize the expected number of wins, so that two policies are equivalent if and only if the expected number of wins obtained by following both is equal:

$$\pi_{\theta_1} \sim_g \pi_{\theta_2} \iff \mathbb{E}_{P_{\pi_{\theta_1}}(h)}(\mathbf{1}_{h_i=1}) = \mathbb{E}_{P_{\pi_{\theta_2}}(h)}(\mathbf{1}_{h_i=1}), \tag{17}$$

where $\mathbf{1}_{h_i=1}$ denotes an indicator function which is one if the ith observation in h is 1 and zero otherwise. Thus, for every $i = 1, ..., N_L^h + N_R^h$, the goal-state S_i is defined as:

$$S_i = \{P(h) :\ \text{s.t. } \mathbb{E}_{P(h)}(\mathbf{1}_{h_i=1}) = j\}, \tag{18}$$

and the "goal-distance" between P_{π_θ} and S_i is given by:

$$D_{KL}(S_i||P_{\pi_\theta}) = \min_{P \in S_i} D_{KL}(P||P_{\pi_\theta}) = \sum_h P_j^\star(h) \log \frac{P_j^\star(h)}{P_{\pi_\theta}(h)}, \tag{19}$$

where $P_j^\star(h)$ is the distribution in S_j closest to $P_\pi(h)$ in the KL sense, as defined in Eq.11 above. Using Eq.16 we can now compute the goal-distance gradient to obtain:

$$\nabla_\theta D_{KL}(S_i||P_{\pi_\theta}) = \frac{\sum_h P_j^\star(h) N_R^h}{1-\theta} - \frac{\sum_h P_j^\star(h) N_L^h}{\theta}, \tag{20}$$

giving the optimal policy parameter:

$$\theta^* = \mathbb{E}_{P_j^\star(h)}\left(\frac{N_L^h}{N_L^h + N_R^h}\right). \tag{21}$$

In words, the policy maximizing the likelihood of reaching goal-state S_i, is one matching the expected choice probability of $P_j^\star(h)$. Interestingly, such "probability-matching" strategies have been reported in empirical choice behavior in iterated binary choice tasks [8].

5.2 The flow of experience - transition sensitive goals

So far, we only considered goals that are sensitive to individual action or observation counts within experience sequences. Indeed, standard reinforcement learning algorithms can be seen as special cases of our framework under a goal of maximizing a (possibly discounted) count of particular observations defined a-priori as rewarding. Real biological agents however, typically pursue more ecological and complex goals, reflecting preferences over higher-order statistics of experience sequences. For example, people engaged in activities such as sports, meditation, artistic creation, game playing or other challenging tasks, often describe the goal of such pursuits as entering a state of "flow"[6], i.e., certain patterns of experience which cannot be reduced to singular actions or outcomes. To explore the potential of our framework, we consider a class of goals defined by maximizing a weighted sum over the number of different possible observation-action and action-observation transitions:

$$g_{\alpha,\beta}(h) = \sum_{i=1}^{|\mathcal{O}|} \sum_{j=1}^{|\mathcal{A}|} (N_{ij}^h \alpha_{ij} + M_{ji}^h \beta_{ji}), \tag{22}$$

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. To make this concrete, we consider a setting where observations depend only on the immediately preceding actions and actions depend only on the immediately preceding observations. 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:

$$P_\pi(h) = e(o_1)\pi(a_1|o_1)\prod_{i=2}^{n} e(o_i|a_{i-1})\pi(a_i|o_i). \tag{23}$$

This distribution can be parameterized using the switching probabilities $e_{ji} \equiv e(o_t = i|a_{t-1} = j)$ (with the 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:

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

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:

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

where we define the distributions π_α and e_β as:

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

and

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

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

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

Thus, for a fixed environment, maximizing goal $g_{\alpha,\beta}(h)$ corresponds to minimizing the cross entropy between the true policy π, and a reference one π_α, induced by the observation-action transition weight parameters α_{ij}. This goal therefore can be viewed as inducing a naturally emergent "information-cost" over policies, extending previous approaches for explaining goal-directed learning in terms of internal complexity costs [2]. Furthermore, by considering goals that are sensitive to higher-order transition statistics, the current approach can be used to uncover preferences over certain flows of experience (action-observation subsequences), beyond mere accumulation of externally-determined "rewarding" outcomes.

References

[1] S. Aberbach-Goodman, B. Buaron, L. Mudrik, and R. Mukamel. Same action, different meaning: neural substrates of action semantic meaning. *Cerebral Cortex*, 32(19):4293–4303, 2022.

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

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

[4] M. Bowling, J. D. Martin, D. Abel, and W. Dabney. Settling the reward hypothesis. *arXiv preprint arXiv:2212.10420*, 2022.

[5] T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[6] M. Csikszentmihalyi and I. S. Csikszentmihalyi. *Optimal experience: Psychological studies of flow in consciousness*. Cambridge university press, 1992.

[7] J. D. Dunne. *Foundations of Dharmakirti's philosophy*. Simon and Schuster, 2004.

[8] I. Erev and G. Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912, 2005.

[9] L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti. Parietal lobe: from action organization to intention understanding. *Science*, 308(5722):662–667, 2005.

[10] D. Hume. *A treatise of human nature*. Clarendon Press, 1896.

[11] L. P. Kaelbling. Learning to achieve goals. In *IJCAI*, volume 2, pages 1094–8. Citeseer, 1993.

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

[13] Y. C. Leong, B. L. Hughes, Y. Wang, and J. Zaki. Neurocomputational mechanisms underlying motivated seeing. *Nature human behaviour*, 3(9):962–973, 2019.

[14] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In *AI&M*, 2006.

[15] M. Minsky. Computation: Finite and infinite machines prentice hall. *Inc., Engelwood Cliffs, NJ*, 1967.

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

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

[18] M. R. Roesch, A. R. Taylor, and G. Schoenbaum. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron*, 51(4):509–520, 2006.

[19] E. Rosch and C. B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605, 1975.

[20] M. Siderits, T. J. Tillemans, and A. Chakrabarti. *Apoha: Buddhist nominalism and human cognition*. Columbia University Press, 2011.

[21] M. Song, Y. K. Takahashi, A. C. Burton, M. R. Roesch, G. Schoenbaum, Y. Niv, and A. J. Langdon. Minimal cross-trial generalization in learning the representation of an odor-guided choice task. *PLoS Computational Biology*, 18(3):e1009897, 2022.

[22] A. Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

[23] L. Wittgenstein. *Philosophical investigations*. Wiley-Blackwell, New York, NY, USA, 1953.