Agent-centric learning: from external reward maximization to internal knowledge curation

Hanqi Zhou[1,2], **Fryderyk Mantiuk**[1], **David G. Nagy**[1,2], **Charley M. Wu**[1,2,3,4]
hanqi.zhou@uni-tuebingen.de

[1]**Human and Machine Cognition Lab, University of Tübingen, Germany**
[2]**Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Germany**
[3]**Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany**
[4]**Hessian.AI, Darmstadt, Germany**

Abstract

The pursuit of general intelligence has traditionally centered on external objectives: an agent's control over its *environments* or mastery of specific *tasks*. This external focus, however, can produce specialized agents that lack adaptability. We propose *representational empowerment*, a new perspective towards a truly agent-centric learning paradigm by moving the locus of control inward. This objective measures an agent's ability to controllably maintain and diversify its own knowledge structures. We posit that the capacity—to shape one's own understanding—is an element for achieving better "preparedness" distinct from direct environmental influence. Focusing on internal representations as the main substrate for computing empowerment offers a new lens through which to design adaptable intelligent systems.

1 The challenges of task- and environment-centric learning

"Intelligence is what you use when you don't know what to do." — Jean Piaget

Reinforcement Learning (RL) has made great progress in training agents to excel at narrow tasks by maximizing rewards (Sutton & Barto, 1998). Yet as termed by Abel et al. (2024), its core dogmas—the reward hypothesis (all goals as reward maximization) and the environment spotlight (focus on modeling environments over agents)—reveal a tension. An agent optimized for a single task reward in a well-defined environment often struggles when new tasks not incentivized by its original training (Ringstrom, 2022; Alet et al., 2020). Given that agents, over a lifetime, will have to learn many aspects of the world, and since we cannot simulate all possible worlds for them to learn in, the current learning paradigm makes it hard to achieve broadly applicable intelligence—a high level of preparedness for unforeseen challenges.

Addressing the limitations suggests a new look beyond purely external task objectives and environmental designs. A promising direction involves a transition from an external task-centric and environment-centric viewpoint to a more internal agent-centric perspective (Singh et al., 2009). This agent-centric view prioritizes the development of internal representations that allow an agent to understand, adapt, and act effectively even when external objectives are novel or underspecified.

The shift naturally raises a further question: by what *principle* should these internal representations be managed to best prepare an agent for future challenges? The information-theoretic concept of *empowerment* (Klyubin et al., 2005; Salge et al., 2014; Lidayan et al., 2025; Mantiuk et al., 2025) offers a promising but underspecified framework. Empowerment quantifies an agent's potential

to influence its future, often measured by the range and controllability of reachable states in the environment. This inherent link to having diverse options makes it a good candidate for formalizing "preparedness". However, existing approaches to empowerment still focus on an agent's control over *external environmental states*. Instead, a truly agent-centric view invites us to redirect this lens inward. If an agent's capacity is shaped by its internal knowledge, then its ability to control and diversify these internal representations could be a more direct route to robust adaptability.

Here, we offer a new perspective towards agent-centric learning. By applying empowerment to the agent's internal representations, instead of asking which state in the external environment one should reach, we ask **what kind of internal representational structures should an agent form and curate to maximize its "preparedness" for a diverse and unpredictable future?** This focus suggests that lasting adaptability may arise more profoundly from an agent's mastery over its own evolving understanding than from its immediate capacity to alter the external world.

1.1 Extrinsic reward maximization for task-centric learning

A key tenet of RL is that "all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward)" (Sutton & Barto, 1998). This statement, coupled with the view of intelligence as primarily goal achievement (McCarthy, 1998), underpins the influential "reward-is-enough" hypothesis: that maximizing reward is a sufficient objective for developing general intelligence (Silver et al., 2021).

In this formulation, the reward signal is often a scalar feedback to guide the agent learning in an *environment*, indicating the desirability of its actions (or states). Traditionally, the reward comes externally from the environment and is predefined by the designer, about a *task* in their mind to be solved. The agent's objective then becomes finding a solution (policy) that maximizes the reward.

However, this *task-centric learning* (when rewards are tied to specific tasks) faces several problems. It is challenging to craft rewards that precisely capture intended goals without incentivizing undesirable "reward hacking" (e.g., a cleaning robot hiding a mess instead of collecting it; Krakovna et al., 2020; Skalse et al., 2022). Even if the reward function is not correlated with any unintended objectives, it may lack the expressive power to represent all desirable orderings over policies or complex goal structures (Abel et al., 2021; Bowling et al., 2023). This can lead to "steady-state type" failures, where agents repeatedly attempt an impossible action (e.g., phasing through a wall) if that naively maximizes a flawed reward signal (e.g., given the goal of going to the next room).

Thus, while effective for achieving specific tasks in well-defined environments, this exclusive reliance on external reward maximization fails to guide the development of the *generalizable internal representations* essential for long-term adaptability in open-ended problems (Hubinger et al., 2019), such as autonomous robots in unpredictable settings or AI systems for creative discovery.

1.2 Intrinsic motivation offers a band-aid fix towards environment-centric learning

The challenge of developing agents that can generalize beyond training tasks led to *intrinsic motivation*—task-agnostic learning objectives that encourage exploration and skill acquisition. These internal rewards, generated by the agent itself, are typically based on principles like curiosity (seeking novelty or surprise; Pathak et al., 2017; Achiam & Sastry, 2017), learning progress (improving a model of the environment; Houthooft et al., 2016), or competence (achieving self-set goals; Colas et al., 2022).

While intrinsic motivation has demonstrably improved learning about the current environment, it remains fundamentally *environment-centric* in two key ways. First, its objectives are inherently tied to the external world. For example, novelty seeking encourages visiting all environmental states, and learning progress drives the formation of a more accurate transition model of the current environment (Modirshanechi et al., 2022). The focus, while broader than a single task, remains on "what is out there to be known or done". This can lead to overfitting to the specifics of the current environment. Second, intrinsic motivation functions by augmenting or replacing the external reward signal. The

agent's learning still aims to maximize a (now potentially composite) scalar reward. While internal representations are learned and refined in this process, they are developed to the extent that they support this reward-seeking behavior, not necessarily because they possess inherent qualities, e.g., high compositionality, that would directly help future learning in different contexts.

Thus, while intrinsic motivation pushes agents beyond myopic exploitation of external rewards and encourages more thorough engagement with their environment, it does not fully address the challenge of building internal cognitive structures designed for lasting adaptability (Abel et al., 2018). The resulting representations, though often richer, are still predominantly shaped by environmental regularities, which may not generalize to new settings.

1.3 Empowerment as a general-purpose objective

Given that existing reward design schemes (both extrinsic and intrinsic) remain tethered to predefined environments, how can an agent develop more general-purpose adaptability? One promising direction lies in identifying *instrumentally convergent goals* (Bostrom, 2012)—rewards that are beneficial across final goals in a wide range of environments and tasks (Omohundro, 2018). Intuitively, for vastly different long-term objectives, from manufacturing paperclips to exploring galaxies, certain intermediate goals are consistently pursued not because they are inherently valuable, but because they are instrumental stepping stones. For example, any society is likely to develop highly efficient energy transportation, e.g., superconducting cables. This is not because conductivity is a terminal goal for society's values, but because energy is a powerful enabler for nearly any non-trivial goal.

The concept of empowerment, first introduced by Klyubin et al. (2005), offers an information-theoretic way to quantify such a general-purpose objective. Empowerment measures an agent's capacity to reliably bring about diverse futures regardless of the final goal. Formally, we can define the traditional, *environment-centric empowerment* (EnvEmp) as the degree of control an agent has over its external environment through the channel capacity (mutual information I) between its sequence of actions $a_{1:T}$ and the resulting environmental state s' from its current state s:

$$\mathrm{EnvEmp}(s) = \max_{\pi(a_{1:T})} I\left(s'; a_{1:T} \mid s\right) \tag{1}$$

High empowerment signifies that the agent's actions can reliably lead to many distinct environmental states, a capacity that serves as a proxy for preparedness. However, environmental empowerment, while a step towards general-purpose objectives, inherits some limitations previously discussed. By focusing on control over externally defined environmental states s and s', it remains susceptible to the "frame-of-reference" problem (Clancey, 2014). If the definition of these states is imposed externally or is not learnable by the agent, then maximizing environmental empowerment might still lead to specialization to the idiosyncrasies of that particular representation (Mantiuk et al., 2025).

This brings us to the central thesis of this paper: General intelligence and adaptability may require shifting the locus of empowerment from the external environment to the agent's own *internal representations*. The critical capacity might not just be to control the world, but to control, adapt, and expand the very way the agent models and understands the world.

2 Agent-centric representational empowerment

2.1 From environmental to representational states

Environmental empowerment has been shown to explain human behavior that we value the potential of diverse options (Brändle et al., 2023; Du et al., 2023), although it is subject to the environmental dynamics (Mantiuk et al., 2025). Here, we define **agent-centric empowerment** on agent's internal representational structures that maximizes its "preparedness", which in this sense, refers to its capacity to generate or reconfigure its knowledge to effectively address unforeseen tasks.

Consider a Minecraft world (Hafner, 2021), where an agent driven by *environmental empowerment* aims to maximize its control over immediate surroundings. This may translate to building a wooden

fortress optimized for the local terrain in a forest. However, this specialization becomes suboptimal if the agent moves to a barren desert. In contrast, an agent driven by *representational empowerment* focuses on the knowledge of building techniques and material properties, instead of investing all resources into one perfect base. When the challenge changes to woodland, skyblock or trial chambers, it can craft new shelters or tools from available items—without needing prior exposure.

Thus, rather than entirely re-defining empowerment, we re-frame the pertinent states involved. Our formulation follows the spirit of the "bitter lesson" (Sutton, 2019): instead of relying on fixed human-specified knowledge, we propose that agents explicitly learn to build their own knowledge by making the dynamics of representation learning the primary target of optimization.

The Minecraft analogy can generalize to abstract knowledge and naturally incorporates the notion of resource constraints (Lieder & Griffiths, 2020). An agent operates under both memory costs, associated with storing its knowledge library Z_k, and computational costs. Representational empowerment helps mediate the trade-off between them. Computational costs can be understood as twofold: the "offline" effort required for the curator to build and maintain an empowered library, and the "online" effort for the executor to adapt that library to solve a new task. By investing more upfront in offline computation to build a library (at a certain memory cost), the agent can amortize future learning, reducing the online computation needed to solve subsequent problems.

2.2 Potential of diverse and controllable representations internally

Let us assume the agent learns over a sequence of tasks. Each task $\tau_k \in \{\tau_1, \tau_2, \ldots\}$ is a Markov Decision Process (MDP) $(\mathcal{S}_k, \mathcal{A}_k, \mathcal{P}_k, R_k)$, with environmental states $\mathcal{S}_k$ (e.g., the observed blocks) and actions $\mathcal{A}_k$ (e.g., chopping a tree). The agent maintains an internal library of representations, Z_k (e.g., design of diamond axe), accumulated from prior tasks $\tau_1, \ldots, \tau_k$. Upon engaging with task τ_{k+1}, it may acquire a new piece of knowledge $\hat{Z}_{k+1}$ (e.g., construction of bamboo planks).

Here we use meta reinforcement learning (Botvinick et al., 2019) to explain the two components in learning: 1) A meta-level *curator* responsible for evolving the internal representational library Z_k, learned from past tasks $\tau_{1:k}$ to maximize its representational empowerment. 2) A task-level *executor* that uses the curated representation Z_k to find solutions for the following specific task τ_{k+1}.

2.2.1 Curator: representational empowerment maximization

We frame the curator's decision-making at the meta level. At each step k, the curator observes a state $s_k^c = (Z_{k-1}, \hat{Z}_k)$—its current library and new knowledge—and selects an integration action $a_k^c \in \mathcal{A}^c$. These actions, e.g., selecting, composing, or pruning, produce the next library, $Z_k = a_k^c(Z_{k-1}, \hat{Z}_k)$. The curator's goal is to maximize an intrinsic reward $r_k^c(Z_k) = \mathrm{RepEmp}(Z_k)$, the *representational empowerment* of the resulting library Z_k (defined below in Eq. 2).

This reward $\mathrm{RepEmp}(Z_k)$ is calculated via multiple simulated roll-outs. The agent imagines applying a sequence of **modification operations**, $\omega_k^{1:T} = \{\omega_k^1, \ldots, \omega_k^T\}$ drawn from a set of available operations Ω, to Z_k. Using $\omega_k^{1:T} = \boldsymbol{\omega}_k$ for simplicity, this yields a modified library $Z'_k \sim p(Z'_k \mid Z_k, \omega_k)$. Then $\mathrm{RepEmp}(Z_k)$ is the channel capacity between these imagined operations and their outcomes, quantifying control over the agent's own representational state:

$$\mathrm{RepEmp}(Z_k) = \max_{\boldsymbol{\omega}_k \in \Omega^T} I(Z'_k; \boldsymbol{\omega}_k \mid Z_k) \tag{2}$$

Here, $\boldsymbol{\omega}_k$ is sampled from the Ω^T, T-fold Cartesian product of Ω, and denotes an operation sequence over T time steps. A high empowerment value, resulting as the reward for action a_k^c, means the library Z_k is both diverse and controllable. The horizon T also reflects a computational budget for how much Z_k can be internally modified by the curator, or by the executor to adapt it later.

We aim to distinguish the high-level actions (a^c), on how to update the library globally, from primitive modification operations (ω). Operations ω are fine-grained transformations to update pieces of

knowledge representation, e.g., the continuous interpolation of high-dimensional features, or symbolic rules (e.g., mutation) for discrete modules, from Z_{k-1} and $\hat{Z}_k$.

Decomposing $\mathrm{RepEmp}(Z_k)$, we can see better that it encourages libraries Z_k that have the potential for diversity (i.e., can be modified into many distinct forms Z'_k) and controllability (i.e., the transformation cannot be arbitrary):

$$I(Z'_k; \omega_k \mid Z_k) = H(Z'_k \mid Z_k) - H(Z'_k \mid Z_k, \omega_k) \tag{3}$$

The first term, $H(Z'_k \mid Z_k)$, quantifies the **diversity** of representation Z'_k reachable from Z_k. Higher diversity suggests Z_k can be transformed into a wide range of different representations, potentially useful for an as-yet-unknown task τ_{n+1}. The second term, $H(Z'_k \mid Z_k, \omega_k)$, quantifies the average **uncertainty** of the outcome Z'_k given a sequence of operations ω_k. A lower value implies that the operations have more predictable effects, enabling precise control. This term favors representations that are not only broadly transformable but also have controllable evolutions.

2.2.2 Executor: task-specific adaptation

Once the curator establishes an empowered representation Z_k, the executor uses it for the next task τ_{k+1} in two potentially intertwined phases:

Representation tuning: the executor may first apply a bounded sequence of operations $\omega_k^{1:T'}$ to mold the generic library into a task-tailored variant $Z_k^\star$. Because Z_k was optimized for high adaptability, a short horizon T' may suffice to reach a configuration beneficial for the new task.

Task completion: Using $Z_k^\star$ (or Z_k directly), the executor interacts with the environment to maximize extrinsic reward R_{k+1}. The executor can also interleave further operations (fine-tuning $Z_k^\star$) with policy updates. For example, if a particular skill from $Z_k^\star$ is almost effective but needs slight adjustments, the executor can adjust it. This creates a *use-improve* cycle: observed task performance provides feedback on which representational refinements are most beneficial, turning the curator's long-term investment into improved sample efficiency and asymptotic performance.

3 Example: empowerment through curating the program library

The representations Z are preferably symbolic (e.g., programs, objects) to support interpretable representational operations, such as abstraction, composition (Rule et al., 2024; Ellis et al., 2021; Zhou et al., 2024). We provide an example of how to instantiate representational empowerment through symbolic programs (Sec. 3.1), which offer several advantages as a representational format (e.g., compositionality and generalization; Lake et al., 2015; Rule et al., 2020).

Formally, we can define a space of program representations $\mathcal{L}$ where each program $z \in \mathcal{L}$ could be a causal model (Icard, 2017), policies (Correa et al., 2025), values, or goals (Davidson et al., 2025) forming the cognitive structure. For generality, each program contains a function term and parameters, e.g., `play(instrument)` has the function `play` and can have parameter `violin` which is already highly abstract with typed parameters (type `instrument`). Drawing inspiration from evolutionary algorithms and genetic programming (Forsyth, 1981), the representational operations Ω can include: **selection** for saving effective programs, function-level **crossover** for combining fragments from multiple programs to create new ones (O'Donnell et al., 2009), **abstraction** for creating higher-level programs (Bowers et al., 2023), and parameter-level **mutation** for modifying parameters of existing programs (Fränken et al., 2022).

3.1 Learning melodic programs

Consider an agent facing a sequence of tasks $\{\tau_1, \tau_2, \ldots\}$. Each task τ_k requires memorizing and playing a specific target melody, M_k^{target}. For a task τ_k, the executor uses the current library Z_{k-1} to match the melody M_k^{target}. Its actions can be primitive (e.g., `add_note(C4)`) or executing a

program (e.g., `repeat(C4, 2)`). The executor receives a reward, R_k, based on the similarity between its generated melody and the target. An empowered library Z_{k-1} would enable generation more efficiently than `add_note` verbatim. When finishing τ_k, the curator might get a new melodic fragment, $\hat{Z}_k$. This new program is a candidate for the library. The curator then selects an integration action a_k^c to produce the new library $Z_k = a_k^c(Z_{k-1}, \hat{Z}_k)$ guided by maximizing $\mathrm{RepEmp}(Z_k)$.

Programs evolution. After exposure to some tasks (e.g., playing simple folk melodies), the library Z_{k-1} contains: `up/down(n, steps)` (increases/decreases the pitch), `repeat(pattern, times)`. In a new task τ_k (e.g., playing chordal harmony and arpeggios), the agent learns two new programs $\hat{Z}_k$, `arpeggio(root, chord, direction)` and `sequence(note, pattern)`. `arpeggio` generates an arpeggio starting from `root`, using notes from the `chord` type (e.g., major, minor). `sequence` generates notes starting from `note`, following the `pattern`.

The agent can apply modification operations on these programs (Z_{k-1} and $\hat{Z}_k$). For example, the agent recognizes that `up` and `down` are special cases of `move(direction, n, steps)` with direction so it **abstracts** over them. The agent can combine, applying **crossover** over `arpeggio` and `repeat` to create `repeated_arpeggio(root, chord, direction, times)`.

3.2 Curating a library with regularized diversity

The agent curates its library Z_k based on representational empowerment (Eq. 2). Rather than maximizing raw diversity ($H(Z'_k \mid Z_k)$ in Eq. 3), the process balances two key principles: task relevance, by integrating useful new programs ($\hat{Z}_k$), and controllable adaptability, by penalizing transformational uncertainty ($H(Z'_k \mid Z_k, \omega_k)$).

Task relevance as a filter. A newly synthesized program $\hat{z} \in \hat{Z}_k$ (e.g., `arpeggio` derived from task τ_5) is considered for long-term integration into Z_{k-1}, because it has proved usefulness within the task τ_k. This task performance acts as an initial, pragmatic filter.

Avoiding a single nearly universal program. This term $H(Z'_k \mid Z_k, \omega_k)$ in Eq. 3, which is subtracted, quantifies the average *ambiguity or lack of precision*. Consider a representation Z_k that is overly flexible, if a program like `generate_any_melody(latent)` exists. It could be a very large, unconstrained, but well-trained neural decoder mapping from a latent space to represent melodies. Here, typical operations (e.g., `mutation` for small parameter perturbations) result in highly wild and unpredictable changes to the melody produced. While it might theoretically be able to produce any melody, $p(Z'_k \mid Z_k, \omega_k)$ is diffuse and high, and the problem for the agent becomes finding a way to choose from the latent space, in order to reliably arrive at the desired melody. Such a representation can prevent the agent from effectively "sculpting" useful representations.

Potentially, after the agent evaluates which representations provide the most empowerment, it decides to keep `move`, `arpeggio`, but discards the now-redundant `up` and `down`.

3.3 Comparing program libraries

We further illustrate how empowerment guides this by comparing potential libraries. For simplicity, we assume two available representational operations Ω: **crossover** and **mutation**.

Diversity preference. Assume two melodic library candidates: $Z_A = \{\texttt{up}, \texttt{down}\}$ and $Z_B = \{\texttt{move}, \texttt{repeat}\}$ (programs explained in Sec. 3.1). Each program can be applied with **crossover** into $M = 3$ distinct variants by converting its style, rhythm, or articulation (e.g., `up_staccato`, `move_smooth`, `repeat_accelerando`).

Modifications for the 2 programs in library Z_A can lead to $M^2 = 9$ distinct libraries Z'_A (e.g., $\{\texttt{up\_staccato}, \texttt{down}, \texttt{repeat}\}$). Though they are syntactically different, the functional diversity might be less. Crossovers of `up` and `down` overlap because of the music octave. We estimate that the 9 variants from `up` and `down` together yield approximately $M + \delta \approx 6$ effectively distinct libraries, $N_{\text{eff}}(Z_A) \approx 6$. So, $\mathrm{RepEmp}(Z_A) = H(Z'_A \mid Z_A) \approx \log_2(6) \approx 2.59$ bits.

The library Z_B also has 9 syntactically distinct libraries Z'_B. Here `move` provides richer functions because **mutation** can create higher-order structures considering parameter `direction`, e.g., `move_staccato_rhythmic(rhythm_X, steps)` (applies a rhythmic pattern to the movement while having staccato). These 3 variants of `move` could provide at least $2 \times M = 6$ distinct functional capabilities. Thus, $N_{\text{eff}}(Z_B)$ is estimated as $6 \times 3 = 18$. So, $\text{RepEmp}(Z_B) = H(Z'_B \mid Z_B) \approx \log_2(18) \approx 4.17$ bits.

Controllability preference. Assume a new library candidate Z_C = {`neural_gen(latent)`, `repeat`}. Here, `neural_gen` is a powerful neural melody generator. For `neural_gen`, suppose there are 20 mutations (the size of the latent space for its parameter). Among them, outcomes of 5 latents are predictable ($\omega^{\texttt{pred}}$) and unique. outcomes of 15 latents are unstable (ω^{unstable}), having equal chance of producing 'style alpha' and 'style beta'. The diversity of $N_{\text{eff}}(Z_B) \approx 7 \times 3 = 21$, and $H(Z'_C \mid Z_C) \approx \log_2(21) \approx 4.39$ bits which reflects high potential. There is a punishment for the uncertainty that the policy considers $20 \times 3 = 60$ combinations, thus $H(Z'_C \mid Z_C, \omega_C) = \frac{(5\times3\times0)+(15\times3\times\log 2\text{ bit})}{60} = \frac{15}{23} \approx 0.75$ bits. So, $\text{RepEmp}(Z_C) \approx 4.39 - 0.75 = 3.64$ bits.

Decision and interpretation. The library Z_B with the more abstract program (`move`) is more empowered, $\text{RepEmp}(Z_B) > \text{RepEmp}(Z_A)$ ($4.17 > 2.59$), because its components can be transformed into a more functionally diverse set. The diversity of $Z_C(4.39)$ is actually higher than Z_B (4.17) because of a powerful general program `neural_gen`. However, when considering the penalization of uncontrollable mutations, i.e., those programs which require another search over parameter space, Z_B is preferred.

4 Discussions

We propose an agent-centric learning paradigm, based on *representational empowerment*, in which an agent maximizes its capacity to controllably diversify its internal knowledge, rather than the external world. This framework offers a new direction for building adaptable agents while also extending key ideas from AI, cognitive science, and evolutionary theory.

Knowledge cultivation in AI. From a continual learning perspective, an agent's internal library Z_k is its evolving knowledge. Unlike standard Bayesian updating (e.g., Bayes-Adaptive MDPs maintaining beliefs over MDPs) that aims to assimilate *all* new evidence $\hat{Z}_k$, representational empowerment guides a *selective* curation of Z_k rather than updating beliefs based on a fixed parameterization of past experiences (Bowling & Elelimy, 2025). The goal of curating a library for fast adaptation is also conceptually related to meta-learning, e.g., Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017). MAML learns a parameter initialization that can be quickly fine-tuned to new tasks via gradient descent. However, MAML's meta-objective is only tied to performance on a distribution of tasks. Representational empowerment, in addition, uses an intrinsic objective—the empowerment of the library itself—to foster task- and environment-agnostic adaptability.

That said, operationalizing $\text{RepEmp}(Z_k)$ (Eq. 2) presents both key challenges but also promising future directions. First, the framework's effectiveness is sensitive to the (pseudo-)metric or kernel used to measure distance in the representation space, which is important for computing entropy terms $H(Z'_k \mid Z_k)$ and $H(Z'_k \mid Z_k, \omega_k)$. A purely syntactic metric might be brittle. A more robust approach could define a *functional* metric, where the "distance" between two representations is measured by the behavioral difference they produce when executed. This metric could even be learned, for instance through contrastive methods, to capture a task-relevant notion of similarity.

Second, representational empowerment is sensitive to the predefined set of modification operations, Ω. One direction is to treat the set of operations not as fixed, but as a dynamic, learnable component. An agent could start with a set of primitive, cognitively-inspired operations (e.g., copy, compose, abstract) and learn to construct more powerful macro-operations over time. This turns Ω into a second-order library that co-evolves with the primary knowledge library Z_k, creating a virtuous cycle of cognitive growth and aligning with the idea of a shared "cognitive toolkit" discussed below.

Information processing in cognitive science. Aligning with the perspective of resource rationality in cognitive science (Lieder & Griffiths, 2020), representational empowerment exchanges computational costs with memory costs brought by the library Z_k, which amortizes the learning effort. This extends with classical information-theoretic objectives that can be agnostic to the cost of deriving or searching for representations. For instance, frameworks like the Information Bottleneck (Tishby et al., 2000), which compresses an input X to preserve information about a *specific target* Y, or Predictive Information (Bialek et al., 2001), which captures past-future regularities, typically compress representations for a predefined data stream for "memory" regardless of any "computation". They are fundamentally about the efficient processing of sensory information. Here, we argue that an agent's long-term cognitive effectiveness may be better understood not just by how efficiently it processes information about its environment, but by the adaptive potential it cultivates within its own representational system.

Socio-cultural dynamics. In a multi-agent context, the set of representational operations Ω can become a shared "cognitive toolkit". Though agents may share the same physical world, their different goals, values, and learned world models mean that the "effective task" each agent faces and the resulting representations are often unique (Molinaro & Collins, 2023; Witt et al., 2024). Two agents in the same physical space can have divergent policies and interpretations of environmental affordances. Consequently, directly transferring policies or internal representations from one agent to another is often difficult, while the ability to exchange or co-develop these *operations* is crucial for collective intelligence, just as human progress leverages shared conceptual tools (e.g., language, mathematics; Wu et al., 2024). For example, effective education often focuses on problem-solving methodologies (e.g., mathematical proof techniques like induction) rather than just rote memorization of solutions to specific problems. Future work could explore how a population can collectively discover and disseminate operations that enhance their collective representational empowerment, forming a "cultural ratchet" for cognitive tools (Tennie et al., 2009).

Acknowledgments

We thank Alison Gopnik, Eunice Yiu and Fei Dai for helpful discussions. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting HZ. HZ, DGN, & CMW are supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645, and funded by the DFG under Germany's Excellence Strategy – EXC 2117 – 422037984.

References

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, pp. 10–19. PMLR, 2018.

David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021.

David Abel, Mark K Ho, and Anna Harutyunyan. Three dogmas of reinforcement learning. *arXiv preprint arXiv:2407.10583*, 2024.

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*, 2017.

Ferran Alet, Martin F Schneider, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Meta-learning curiosity algorithms. *arXiv preprint arXiv:2003.05325*, 2020.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85, 2012.

Matthew Botvinick, Sam Ritter, Jane X Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in cognitive sciences*, 23(5):408–422, 2019.

Matthew Bowers, Theo X Olausson, Lionel Wong, Gabriel Grand, Joshua B Tenenbaum, Kevin Ellis, and Armando Solar-Lezama. Top-down synthesis for library learning. *Proceedings of the ACM on Programming Languages*, 7(POPL):1182–1213, 2023.

Michael Bowling and Esraa Elelimy. Rethinking the foundations for continual reinforcement learning. *arXiv preprint arXiv:2504.08161*, 2025.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In *International Conference on Machine Learning*, pp. 3003–3020. PMLR, 2023.

Franziska Brändle, Lena J Stocks, Joshua B Tenenbaum, Samuel J Gershman, and Eric Schulz. Empowerment contributes to exploration behaviour in a creative video game. *Nature Human Behaviour*, 7(9):1481–1489, 2023.

William J Clancey. The frame of reference problem in the design of intelligent machines. In *Architectures for intelligence*, pp. 357–423. Psychology Press, 2014.

Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *Journal of Artificial Intelligence Research*, 74:1159–1199, 2022.

Carlos G Correa, Sophia Sanborn, Mark K Ho, Frederick Callaway, Nathaniel D Daw, and Thomas L Griffiths. Exploring the hierarchical structure of human plans via program generation. *Cognition*, 255:105990, 2025.

Guy Davidson, Graham Todd, Julian Togelius, Todd M Gureckis, and Brenden M Lake. Goals as reward-producing programs. *Nature Machine Intelligence*, 7(2):205–220, 2025.

Yuqing Du, Eliza Kosoy, Alyssa Dayan, Maria Rufova, Pieter Abbeel, and Alison Gopnik. What can ai learn from human exploration? intrinsically-motivated humans and agents in open-world exploration. In *Neurips 2023 workshop: Information-theoretic principles in cognitive systems*, 2023.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd acm sigplan international conference on programming language design and implementation*, pp. 835–850, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Richard Forsyth. Beagle—a darwinian approach to pattern recognition. *Kybernetes*, 10(3):159–166, 1981.

Jan-Philipp Fränken, Nikos C Theodoropoulos, and Neil R Bramley. Algorithms of adaptation in inductive inference. *Cognitive Psychology*, 137:101506, 2022.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. *arXiv preprint arXiv:2109.06780*, 2021.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Thomas F Icard. From programs to causal models. In *Proceedings of the Amsterdam Colloquium*, pp. 35–44, 2017.

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.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 3, 2020.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Aly Lidayan, Yuqing Du, Eliza Kosoy, Maria Rufova, Pieter Abbeel, and Alison Gopnik. Intrinsically-motivated humans and agents in open-world exploration. *arXiv preprint arXiv:2503.23631*, 2025.

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

Fryderyk Mantiuk, Hanqi Zhou, and Charley M Wu. From curiosity to competence: How world models interact with the dynamics of exploration. In A Ruggeri, D Barner, C Walker, and N Bramley (eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*, San Francisco, CA, 2025. Cognitive Science Society.

Michael McCarthy. *Spoken language and applied linguistics*. Cambridge University Press, 1998.

Alireza Modirshanechi, Johanni Brea, and Wulfram Gerstner. A taxonomy of surprise definitions. *Journal of mathematical psychology*, 110:102712, 2022.

Gaia Molinaro and Anne GE Collins. A goal-centric outlook on learning. *Trends in Cognitive Sciences*, 27(12):1150–1164, 2023.

Timothy J O'Donnell, Joshua B Tenenbaum, and Noah D Goodman. Fragment grammars: Exploring computation and reuse in language. 2009.

Stephen M Omohundro. The basic ai drives. In *Artificial intelligence safety and security*, pp. 47–55. Chapman and Hall/CRC, 2018.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Thomas J Ringstrom. Reward is not necessary: How to create a modular & compositional self-preserving agent for life-long learning. *arXiv preprint arXiv:2211.10851*, 2022.

Joshua S Rule, Joshua B Tenenbaum, and Steven T Piantadosi. The child as hacker. *Trends in cognitive sciences*, 24(11):900–915, 2020.

Joshua S Rule, Steven T Piantadosi, Andrew Cropper, Kevin Ellis, Maxwell Nye, and Joshua B Tenenbaum. Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications*, 15(1):6847, 2024.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Empowerment–an introduction. *Guided Self-Organization: Inception*, pp. 67–114, 2014.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial intelligence*, 299:103535, 2021.

Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society*, pp. 2601–2606. Cognitive Science Society, 2009.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

Claudio Tennie, Josep Call, and Michael Tomasello. Ratcheting up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364 (1528):2405–2415, 2009.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Alexandra Witt, Wataru Toyokawa, Kevin N Lala, Wolfgang Gaissmaier, and Charley M Wu. Humans flexibly integrate social information despite interindividual differences in reward. *Proceedings of the National Academy of Sciences*, 121(39):e2404928121, 2024. DOI: 10.1073/pnas.2404928121.

Charley M Wu, Rick Dale, and Robert D Hawkins. Group coordination catalyzes individual and cultural intelligence. *Open Mind*, 8:1037–1057, 2024. DOI: 10.1162/opmi_a_00155.

Hanqi Zhou, David G Nagy, and Charley M Wu. Harmonizing program induction with rate-distortion theory. *arXiv preprint arXiv:2405.05294*, 2024.