Artificial Agency Program:
Curiosity, compression, and communication in agents

Richard Csaky
richard.csaky@gmail.com

Abstract

This paper presents the *Artificial Agency Program* (AAP), a position and research agenda for building AI systems as reality embedded, resource-bounded agents whose development is driven by curiosity-as-learning-progress under physical and computational constraints. The central thesis is that AI is most useful when treated as part of an extended human–tool system that increases sensing, understanding, and actuation capability while reducing friction at the interface between people, tools, and environments. The agenda unifies predictive compression, intrinsic motivation, empowerment and control, interface quality (*unification*), and language/self-communication as selective information bottlenecks. We formulate these ideas as a falsifiable program with explicit costs, staged experiments, and a concrete multimodal tokenized testbed in which an agent allocates limited budget among observation, action, and deliberation. The aim is to provide a conceptual and experimental framework that connects intrinsic motivation, information theory, thermodynamics, bounded rationality, and modern reasoning systems[1].

1 Introduction

Contemporary frontier AI systems are increasingly capable, yet the dominant training and evaluation pipelines still under-emphasize the conditions under which biological agency develops: reality embedded interaction, finite compute and memory, constrained sensing and actuation, and the continual need to act under uncertainty (Sutton and Barto, 2018; Schmidhuber, 2010; Oudeyer et al., 2007; Friston, 2010; Lake et al., 2017; LeCun, 2022). We start from a practical thesis: AI should be designed as part of a larger human–tool system that improves sensing, understanding, and actuation under realistic constraints, while remaining efficient and interpretable.

This framing emphasizes *agency gain* and *interface friction* simultaneously. Increasing raw capability alone is not sufficient if the resulting system is hard to steer, poorly coupled to human intent, or unable to operate efficiently under the time, energy, and communication budgets that dominate real environments. In many settings, the bottleneck is not whether a model can in principle provide a solution, but whether the coupled human–AI system can reliably gather information (observations), allocate limited compute, communicate intent, and act in an effective way. We therefore treat agency as a property of the coupled system, not just the standalone model (Clark and Chalmers, 1998; Lieder and Griffiths, 2020). The strive for ever higher agency is fundamental, as it is a direct consequence of human nature defined by curiosity and playfulness aimed at discovering—predicting—the world.

A seminal work of John Archibald Wheeler postulates that physics exists due to there being someone to observe it, more precisely to ask yes/no questions (Wheeler, 1989). This motivates the curiosity

[1]This is a working draft. Feedback and criticism is most welcome.

objective presented in Schmidhuber (2010)—increasing the rate of compression by seeking patterns (observations) on which prediction improvement is maximal given current capabilities. This avoids completely random patterns or patterns which are easy to compress and is analogous to the rate at which we can ask-and-answer questions about the universe, i.e. spatiotemporal patterns. Therefore we take this as a single core objective of human and artificial agents.

1.1 Human frames of reference and extended agency

Many failure modes of current systems are difficult for people to anticipate precisely because the systems are trained and deployed under non-human conditions. Internet-scale next-token training with effectively superhuman memory for textual regularities and weak grounding in action can produce competent behavior, but it does not reproduce the developmental constraints that shaped human cognition. Human intelligence is structured by limited memory, limited sensing and actuation bandwidth, limited processing speed, partial observability, and the need to rely on tools and other people for increasing agency and cognitive offloading (Clark and Chalmers, 1998; Ryan and Deci, 2000; Lake et al., 2017). Therefore parts (or dimensions) of human cognition must necessarily arise precisely due to the restrictions imposed by our specific biological instantiation from a more general intelligence manifold.

This suggests a distinction between *capability* and *distance from human constraints*. Two systems may achieve similar performance (intelligence) while occupying different regions of a constraint manifold defined by sensing modalities, action affordances, energetic pressure, temporal scale, and memory budget. We propose measuring both overall competence (e.g., predictive performance, empowerment, control) and constraint proximity, because these jointly affect interpretability, collaboration quality, and failure modes (Chollet, 2019; Lieder and Griffiths, 2020). By representing intelligence as such a manifold, one may quantify overall competence as the magnitude of a point on the manifold, providing a scalar comparison between human and artificial intelligence. Human-likeness (or alienness) can be measured with a distance metric between human and AI points on the manifold.

The reason it is hard to encode into AI the fundamental aspects of human intelligence is that they are part of a fundamentally human frame of reference. What is curious or novel to us is defined by our scale in time and space, by evolutionary programming, by biological considerations, by the properties of our sensing and actuating mechanisms. Indeed even by our level of intelligence or cognition—testament to this is the fact that looking at the arch of human (and animal) history, curiosity and inquisitiveness has always shifted—adjusted—depending on our cognitive abilities. A human 1000 years ago could hardly be curious about differential topology. An agent can only be curious with regards to what it can sense, understand and act upon—where all three include tools forming part of the extended self. It is erroneous to simply jump-start intelligence without considering the manifold of constraints it is embedded in.

1.2 Curiosity, compression, and progressive capability expansion

We adopt a learning-progress view of curiosity: intrinsically valuable patterns are neither random nor trivial, but those for which the agent can currently improve compression or prediction (Schmidhuber, 2010; Oudeyer et al., 2007). This makes curiosity inherently capability-relative. An agent can only be curious about patterns it can (in some sense) sense, model, or act upon, and the difficulty of useful tasks should grow with the agent's capabilities rather than being fixed in advance. This view aligns with developmental perspectives in which exploration and play are active processes of skill

acquisition rather than passive consumption of novelty (Ryan and Deci, 2000; Oudeyer et al., 2007).

On this view, curiosity creates pressure for *capability expansion*. If learning progress is limited by sensing, actuation, memory, or compute, the agent should prefer interventions that relax whichever bottleneck most improves future learning progress, subject to cost. This links curiosity to empowerment, control, and communication: an agent may benefit from seeing more, doing more, delegating more, or coordinating with others and tools more effectively. Because the world presents patterns over space and time, this pressure naturally favors longer-horizon prediction and increasingly structured internal models, not merely myopic novelty seeking (Schmidhuber, 2010; Ha and Schmidhuber, 2018; Hafner et al., 2023). Due to this pressure of predicting as far as possible into the future, the study of mathematics, physics, information, etc., should naturally emerge from such intrinsically motivated agents, as they aim to compress spatiotemporal observations into as small a description as possible. The environment is crucial: to develop truly capable agents we need them to be embedded in the physical spatiotemporal reality, not in a computer sandbox. The learning trajectory and actions at any step are constrained by the physical sensing, actuating, and energy resources of the agent.

For example, an agent living in a 2D world may be "frustrated" (and therefore curious) that it cannot "see" a complete circle, only parts of it. Similarly as a human I may be annoyed that I cannot see 3D objects in their entirety at the same time, whereas I can see 2D shapes wholly. The implications are: 1. Since I am currently seeing the world in 3D I may have a desire to see 4D objects, but my desire to see 10D objects is much lower, and 2. I cannot predict the kinds of goals or desires that seeing in 4D (i.e., gaining more sensing capability) would unlock. A similar analogy can be made with respect to actuation and cognition. Perhaps an agent that is aware of its limitations will be most curious about what happens when removing some of those limitations, since this is a future fundamentally harder to predict than asking questions and answering within its current limitations. It is a sort of meta-objective: I not only want to predict the world as I can interact with it now, but I want to predict the kinds of questions I may ask when my capabilities increase. Analogously, my true underlying objective is to predict the hidden (unobservable) state of the environment in its totality.

1.3 Energy–performance frontiers

Efficiency is a natural concern in AAP. Many AI systems become practically useful only when they solve tasks within meaningful time and energy budgets. The guiding intuition is closer to engineering flight than to copying bird biomechanics: progress comes from discovering designs that solve the task under constraints. This motivates explicit energy/compute–performance frontiers and MDL-style criteria for how much task performance is obtained per unit cost (Rissanen, 1978; Grünwald, 2007; Hutter, 2005; Patterson et al., 2021).

We therefore emphasize not only peak performance, but adaptive allocation of computation and communication. In many tasks, the optimal policy is not to spend maximal compute at every step, but to modulate effort based on uncertainty, expected value of information, and opportunity cost (Graves, 2016; Banino et al., 2021; Callaway et al., 2018; Lieder and Griffiths, 2020). This efficiency emphasis also motivates the later focus on private deliberation tokens, action tokens, and observation tokens as interchangeable budgeted resources.

2 Formal Setup

We model an embedded agent interacting with an environment as a partially observed controlled process. Let X_t denote the (generally hidden) environment state, S_t the agent internal state (memory + policy state (θ_t) + latent world-model state), O_t the observation, and $A_t := (U_t, V_t)$ the action.

$$O_t \sim p_O(\cdot \mid X_t;\, c_t^O), \tag{1}$$
$$S_t \sim p_S(\cdot \mid S_{t-1}, O_t, U_{t-1}, V_{t-1};\, c_t^C), \tag{2}$$
$$(U_t, V_t) \sim \pi(\cdot \mid S_t;\, c_t^A), \tag{3}$$
$$X_{t+1} \sim p_X(\cdot \mid X_t, U_t), \tag{4}$$

where c_t^O, c_t^A, c_t^C denote observation, actuation, and compute/interface capacity constraints (e.g., sensor bandwidth, action alphabet/determinism, context length, update budget). These capacities themselves may be modeled as dynamic and agent-controllable.

$$c_{t+1}^O \sim p_{c^O}(\cdot \mid c_t^O, V_t), \tag{5}$$
$$c_{t+1}^A \sim p_{c^A}(\cdot \mid c_t^A, V_t), \tag{6}$$
$$c_{t+1}^C \sim p_{c^C}(\cdot \mid c_t^C, V_t). \tag{7}$$

We view p_S as the agent-side update operator and allow it to include both recurrent/memory updates and (optionally) learning updates of parameters θ_t (e.g., a bounded number of gradient/Bayes-style update steps), all priced under the compute/interface budget c^C.

Learning-progress intrinsic reward. We adopt a Schmidhuber-style learning-progress signal, not raw novelty (Schmidhuber, 2010). Let $\mathcal{L}_{\text{pred}}$ be a predictive loss for future observations over horizon H. For any integers $a < b$, define the interaction segment

$$\mathcal{D}_{a:b} \;:=\; (O_a, A_a, O_{a+1}, A_{a+1}, \ldots, O_{b-1}, A_{b-1}, O_b), \tag{8}$$

where $A_t = (U_t, V_t)$ as in the formal setup.

Let p_θ denote the agent's predictive model for observations, which maps a history to a distribution over the next observation:

$$p_\theta(O_{t+1} \mid \mathcal{D}_{a:t}) \;\equiv\; p_\theta(O_{t+1} \mid O_{a:t}, A_{a:t})\,. \tag{9}$$

We instantiate the horizon-H predictive loss as the cumulative negative log-likelihood

$$\mathcal{L}_{\text{pred}}(\theta; \mathcal{D}_{t:t+H}) \;:=\; \sum_{h=1}^{H} \ell(p_\theta(\cdot \mid O_{t:t+h-1}, A_{t:t+h-1}),\, O_{t+h})\,, \qquad \ell(p, o) := -\log p(o). \tag{10}$$

A generic intrinsic reward is then

$$r_t \;:=\; \mathcal{L}_{\text{pred}}(\theta_{t-H-1}; \mathcal{D}_{t-H:t}) \;-\; \mathcal{L}_{\text{pred}}(\theta_{t-H}; \mathcal{D}_{t-H:t}), \qquad t > H, \tag{11}$$

where $r_t := 0$ for $t \leq H$. This rewards *improvement* in predictive compression rather than complexity or surprise alone, consistent with learning-progress formulations and modern world-model practice (Schmidhuber, 2010; Ha and Schmidhuber, 2018; Hafner et al., 2023).

We combine intrinsic reward with costs:

$$J(\pi, p_S) \;=\; \mathbb{E}\Bigg[\sum_{t=1}^{T} \gamma^{t-1}\Big(r_t - \lambda_O C_O(t) - \lambda_E C_E(t) - \lambda_C C_C(t) - \lambda_M C_M(t)\Big)\Bigg]. \tag{12}$$

Where:

- $C_O(t)$: observation-processing cost (e.g., number of observation tokens/bits ingested or sensor bandwidth used at time t),
- $C_E(t)$: actuation/maintenance cost (e.g., environment-defined energy penalty for executing U_t),
- $C_C(t)$: compute/deliberation/learning cost (e.g., FLOPs, number of deliberation tokens, or number of update steps executed at time t)
- $C_M(t)$ a memory-maintenance cost (motivated by finite-state maintenance and thermodynamic viewpoints on information processing (Still et al., 2012; Parr et al., 2022)).

2.1 Control, empowerment, and plasticity

AAP's hypotheses relate predictive compression to control and empowerment.

Empowerment (actuation capacity into future observations). A standard k-step empowerment definition is the channel capacity from action sequences to future observations (Klyubin et al., 2005a,b; Salge et al., 2014):

$$\mathcal{E}_k(s_t) \;=\; \max_{p(a_{t:t+k-1})} \mathrm{I}(A_{t:t+k-1}; O_{t+k} \,|\, S_t = s_t). \tag{13}$$

Plasticity (observation-to-action influence). We emphasize "plasticity" as environment pressure on the agent. A natural proxy is directed information from observations to actions (Massey, 1990; Kramer, 1998; Abel et al., 2025):

$$\mathcal{P}_{t:T} \;:=\; \mathrm{I}^{\rightarrow}(O_{t:T} \to A_{t:T}) \;=\; \sum_{\tau=t}^{T} \mathrm{I}(O_{t:\tau}; A_\tau \mid A_{t:\tau-1}). \tag{14}$$

High $\mathcal{P}_{t:T}$ can reflect beneficial adaptivity/reactivity. To interpret it specifically as *costly learning pressure* (as opposed to purely reactive control), we report it alongside an explicit update proxy such as $\|\theta_t - \theta_{t-1}\|$ or the number of learning/update steps executed under the compute budget.

Predictive sufficiency / compression. We distinguish: (i) predictive performance on future observations, (ii) compression of hidden state X_t (typically only approximable), and (iii) model complexity / description length of the internal state or policy (Rissanen, 1978; Grünwald, 2007; Tishby et al., 1999; Cover and Thomas, 2006; Bengio et al., 2013). This avoids collapsing all notions of "understanding" into a single score.

2.2 Unification as interface quality

We introduce *unification* as reduction of sensing/acting bottlenecks between agent and environment. We formalize unification as an interface-quality concept.

Let b_O, b_A denote observation and action bottleneck parameters (coarsening, noise, latency, cardinality constraints, etc.). These bottleneck parameters are simply an alternative parameterization of the capacity constraints defined earlier: b_O indexes the observation channel family $p_O(\cdot \mid X; b_O)$ with effective capacity $c^O(b_O)$, and b_A indexes the actuation/policy interface with effective capacity $c^A(b_A)$. We likewise let b_L parameterize a communication/self-communication interface whose cost is accounted for under the same budgeting terms already present in equation (12). Define a task-relative *unification score*

$$\mathcal{U}_t \;:=\; w_O \underbrace{\Big(1 - \frac{\mathcal{R}_O(b_O)}{\mathcal{R}_O^{\max}}\Big)}_{\text{observation losslessness proxy}} + w_A \underbrace{\Big(1 - \frac{\mathcal{R}_A(b_A)}{\mathcal{R}_A^{\max}}\Big)}_{\text{action authority proxy}} + w_L \underbrace{\Big(1 - \frac{\mathcal{R}_L(b_L)}{\mathcal{R}_L^{\max}}\Big)}_{\text{communication bottleneck proxy}}, \tag{15}$$

$$w_O + w_A + w_L = 1, \qquad \mathcal{R}_\cdot \;\geq 0 \tag{16}$$

where $\mathcal{R}_\cdot$ are calibrated inefficiency measures induced by each bottleneck. The exact choice of $\mathcal{R}$ is environment-specific. This way we can test whether improved agents systematically spend resources to increase effective interface quality when allowed to do so.

One way to define the observation-interface inefficiency is equivocation:

$$\mathcal{R}_O \;:=\; H(X|O), \qquad \mathcal{R}_O^{\max} = H(X). \tag{17}$$

$\mathcal{R}_A$ and $\mathcal{R}_L$ can be similarly defined.

3 Hypotheses

3.1 H1: Pragmatic alignment of objectives

In resource-bounded embedded agents, interventions that increase learning progress on future observations also tend to increase useful control over task-relevant environmental degrees of freedom, and vice versa, over a substantial regime of tasks and constraints. Moreover, optimizing equation (12) with respect to agent-side observations implies similar learning progress (or control) over the true environment hidden state from which observations are drawn, when interfaces and bottlenecks are flexible enough to allow for losslessness (i.e. invertibility) in the limit.

This is weaker than strict equivalence between predictive compression, hidden-state compression, and control. It predicts alignment over a regime, not identity for all environments. It is compatible with counterexamples where prediction is easy but uncontrollable, or control is high but not informative.

3.2 H2: Boundary pressure toward unification

When an agent can invest resources to modify its sensing/acting/communication interfaces, optimization of equation (12) will allocate resources toward interface improvements that increase long-horizon learning progress and control, yielding monotonic gains in task-relative unification $\mathcal{U}$

until costs dominate. In finite environments and with favorable scaling of interface cost, a stronger boundary-collapse regime may emerge asymptotically. In an idealized limit where (i) environment complexity is finite and (ii) marginal interface-improvement costs scale favorably, the optimizer can drive $\mathcal{U} \to 1$ (maximal effective interface quality). This as a limit-case conjecture about the objective's asymptotics, not a practical prediction for real systems.

3.3 H3: Constraint-induced predictive/control pressure

Under continued viability constraints (cannot stop interacting), coarse/noisy interfaces, and nonzero costs for action, memory maintenance, and frequent policy updates, agents are driven toward better prediction and selective control because these reduce costly plasticity and wasted computation. Therefore even without directly providing the intrinsic reward of equation (11), an agent may nevertheless be forced to optimize toward it in realistic environments.

This hypothesis aligns naturally with information-theoretic and thermodynamic arguments linking predictive state representations to energetic efficiency (Still et al., 2012; Friston, 2010; Parr et al., 2022; Ortega and Braun, 2013).

3.4 H4: Adaptive compute optimality

We further predict that agent performance under fixed budgets depends not only on model capacity, but on how computation is allocated over time. A meta-controller that dynamically allocates observation, action, and deliberation compute should outperform fixed schedules under the same total budget, especially when task difficulty and observability vary across timesteps. This connects to adaptive computation and metareasoning (Graves, 2016; Banino et al., 2021; Callaway et al., 2018; Lieder and Griffiths, 2020).

3.5 H5: Self-communication bottleneck

An explicit self-communication channel (e.g., text-like or symbol-like private tokens) can improve performance and sample efficiency on tasks requiring long-horizon credit assignment, compositional planning, or coordination, beyond what latent recurrence/planning alone achieves, provided the private channel is bandwidth-regularized. Importantly, the claim is not that verbalized traces are always optimal, but that *selective* self-communication may be an efficient option in a broader action space that also includes direct action and latent deliberation (Yao et al., 2023; Kim et al., 2025; Xie et al., 2025; Wang et al., 2026).

4 Language as an Information Bottleneck

We view language as a selective, lossy, resource-constrained communication channel whose value depends on the environment and on the presence of other agents/tools.

This perspective is consistent with extended cognition (Clark and Chalmers, 1998), communication under resource rationality (Lieder and Griffiths, 2020), and recent reasoning systems that separate or modulate explicit verbal traces from latent computation (Graves, 2016; Banino et al., 2021; Wang et al., 2026). It also helps separate two questions that are often conflated in language-model practice:

Table 1: Hypotheses with positive predictions and disconfirmation paths.

Hyp.	Positive evidence	Disconfirmation / boundary case
H1	Across tasks, interventions improving learning progress also improve control/empowerment metrics under fixed budgets.	Predictive gains with no corresponding control gain on key tasks; or control gains via exploitation that degrade predictive competence.
H2	Agent spends budget to widen sensors/actions/communication only when long-horizon return improves; $\mathcal{U}$ rises then saturates.	Agent never invests in interface improvements despite clear gains; or always over-invests regardless of cost.
H3	Under stronger cost/viability constraints, policies shift toward better predictive state and reduced costly reactive plasticity.	Constraint increases merely collapse behavior without improved predictive organization or selective control.
H4	Meta-control beats fixed observe/act/deliberate schedules at equal budget.	No gain over tuned static schedules, implying control overhead dominates.
H5	Private self-communication tokens improve task performance / sample efficienc.	Private channel collapses into redundant verbosity, or latent-only recurrence dominates under equal compute.

whether language is useful as a training substrate for acquiring broad abstractions, and whether language is the best *online* medium for every intermediate computation once an agent is acting in a world.

From the perspective of a single embedded agent, language is one possible compression channel among many. Its value increases sharply when the environment contains other agents, humans, and tools whose behavior can be altered through communication. In such settings, communication is simply another action class whose utility depends on whether the expected future gain in prediction, control, or coordination exceeds its cost in time, bandwidth, and energy. We therefore treat communication as an instrumentally selected action, not a universally privileged mode.

We also takes seriously the possibility that language-like communication can be useful *internally*. A compact, lossy self-directed channel may support deliberation by forcing the system to encode intermediate structure in a form that is easier to re-read, monitor, and compose. One plausible mechanism is a communicator–interpreter loop: to produce useful messages (even to itself), the agent must model how those messages will be parsed, which can induce a partial third-person stance on its own intermediate states. This motivates concrete experiments comparing latent-only recurrence with explicit private token channels (i.e. standard reasoning traces) under matched cost.

4.1 Implications for LLMs

All LLM-style (including VLMs and VLAs) models should have reasoning modules after having "learned" language. The purpose of a reasoning module however is not to produce a stream-of-words, but rather to allow for deliberation without action. Therefore when a model is post-trained for reasoning on various tasks (including language itself) it should be able to produce sequential hidden state updates without the use of language tokens. Such a model would have the flexibility to choose when to produce and when to avoid self-communication through language and do pure deliberation (without tokenization bottlenecks) instead. Additionally, deliberation can be expanded to any token modality, including vision and action tokens. This is analogous to visual or motor imagery in

neuroscience.

Such a model might observe various types of input tokens and be able to produce and interleave these at any step. Importantly during production there is no fixed order, similar to how a human might pause mid-sentence to do a tiny deliberation loop (with or without explicit language-thoughts). Allowing the model to flexibly start "thinking" while it has not finished a paragraph, or continue thinking while producing an output could help with error recovery and self-monitoring.

In the most general case we propose a modality-agnostic token taxonomy with three roles per modality $m \in \{V, S, T, A, \ldots\}$:

$$m_i \text{ (input)}, \qquad m_s \text{ (private/self)}, \qquad m_o \text{ (public/output)}.$$

This yields a unified interface where the policy can interleave incoming observations, external actions, and internal deliberation:

$$\underbrace{V_i, S_i, T_i}_{\text{incoming}} \Leftrightarrow \underbrace{H, V_s, S_s, T_s, A_s}_{\text{deliberation / imagery / reasoning}} \Leftrightarrow \underbrace{V_o, S_o, T_o, A_o}_{\text{communication / output}} ,$$

where V denotes vision, S denotes audio, T denotes text, and A denotes action tokens. H refers to the output hidden state. The key design question is not whether to allow private tokens, but *when they are worth emitting* under budget, and *which modality is most effective* for deliberation at any given step. This directly supports our compute-efficiency criterion.

It is an open and difficult question how to train such a flexible model given current LLM paradigms. One way to train it for hidden-state deliberation would be to choose blanks (i.e. reserved non-language tokens) whenever the confidence in the next text token is below a threshold. For example one can take a reasoning trace written by a human and run it through a base LLM to get confidence scores for each token. Any position whose score is below a threshold gets shifted to the right and a "blank" is inserted. Then the LLM can be fine-tuned on these augmented traces for a while, and subsequently use the fine-tuned model to regenerate scores for our augmented traces, inserting new blanks according to confidence scores, and so on. After each step the number of positions where blanks are included should decrease and perhaps there is an optimum where no more blanks are required. We should expect the final trace to have many blanks in the beginning and at conceptual jumps in the reasoning. This procedure is still somewhat artificial and not necessarily in line with the goal of using blanks for truly free deliberation. Ultimately, choosing between blanks and other types of imagery modalities (including text) should not be based on language modeling score.

5 Experimental Agenda

AAP is intended to be tested in stages, starting with environments where observability, controllability, and cost structure are explicit, and then moving toward richer interactive multimodal settings. The objective is not only to maximize a given benchmark score, but to validate the proposed metrics, identify counterexamples, and characterize the regimes in which the hypotheses hold.

5.1 Stage 1: Synthetic POMDPs

The first stage should use toy but diagnostic partially observed environments with known latent dynamics and controllable bottlenecks. Gridworld-like domains with sensor coarsening, latency,

observation noise, and switchable action-set cardinality are particularly useful because they permit direct intervention on the variables that define AAP's unification and cost definitions. This stage is primarily for calibrating the metrics, stress-testing H1–H3, and identifying degenerate regimes (e.g., environments where prediction is cheap but irrelevant, or control is high but uninformative).

5.2 Stage 2: ARC-AGI-style interactive inference

ARC-AGI-3-style tasks are attractive because they emphasize compositional priors and generalization under sparse data (Chollet, 2019). Here we can directly test interactive perception–action–deliberation with explicit compute and self-communication costs. This shift makes it possible to study when additional observation is worth requesting, when internal computation is worth spending, and when direct action is preferable. It also creates a clean setting for evaluating self-communication channels.

5.3 Stage 3: Multimodal VLA meta-control

The third stage introduces a pretrained multimodal model as a perception and world-model backbone, together with a lightweight meta-controller that decides at each step whether to acquire more input, act on the environment, or spend budget on private deliberation and adaptation. A frozen backbone plus lightweight adaptation (e.g., LoRA, recurrent adapters, or external memory) is a practical starting point; richer online learning can be introduced later as compute allows. This stage makes the AAP hypotheses concrete in a setting that is close to current multimodal systems.

5.4 Efficiency and pareto optimality

A useful diagnostic is an energy/compute–performance frontier. Let P be task performance and C a normalized cost. We can summarize each policy by (C, P) and compare to an empirical frontier $\mathcal{F}$. One scalar score is L2 distance to the frontier:

$$d_{\mathcal{F}}(C, P) \;=\; \inf_{(c,p)\in\mathcal{F}} \big\|(C, P) - (c, p)\big\|_2, \tag{18}$$

or a task-weighted regret relative to frontier points at matched cost. This formalizes one of our core intuitions: intelligent systems should allocate compute adaptively, not just maximize score irrespective of expenditure.

Importantly in the case of large, pre-trained models the cost should include not just inference but the total pre-training energy usage. Ideally AI systems should be able to dynamically adjust energy usage in a pareto-optimal way, i.e. a particular system's distance from the pareto-optimal curve is the same at any point. A SOTA system then should provide an upper envelope on all previous models across varying levels of energy usage.

5.5 Proof of concept study

A minimal first study can be executed with modest resources by using an ARC-AGI-3-like grid environment with a hidden state and patchwise observations. The agent interacts through three meta-actions—`observe`, `act`, and `deliberate`—and each emitted or consumed token incurs an

explicit cost. Observation and action channels can be bottlenecked by cardinality limits, noise, cropping, or latency, allowing direct manipulation of the quantities used in the unification proxy.

For the model backbone, a compact transformer with image-token and text-token support is sufficient in the first instance. The meta-controller can be a lightweight sequence model over the three meta-actions, trained with a policy-gradient or bandit-style objective on equation (12). A practical initialization is to keep the backbone frozen, treat private deliberation as either latent recurrence or private tokens, and permit sparse low-rank adaptation only after the meta-controller behavior stabilizes. Additional actions aimed at modifying interface bottlenecks can be included.

The key comparisons are straightforward: adaptive meta-control versus fixed observe/act/deliberate schedules; latent-only recurrence versus explicit private tokens; and tight versus loose observation/action bottlenecks, each with and without deliberation cost penalties. Success should not be defined solely by raw task score. The primary criteria are frontier improvement at matched cost, nontrivial and interpretable use of private computation, and systematic shifts in meta-action allocation as interface bottlenecks and costs change. These outcomes would provide direct evidence about H4–H5 while also producing diagnostic signals for H1–H3.

6 Related Work

AAP is most directly aligned with the intrinsic-motivation and developmental-robotics literature, especially work that treats curiosity as *learning progress* rather than raw novelty or surprise (Schmidhuber, 2010; Oudeyer et al., 2007). This distinction is central to the program: novelty-only signals invite stochasticity seeking, whereas learning-progress signals favor patterns that are currently learnable but not yet mastered. Modern exploration methods based on prediction error or novelty remain highly relevant as practical baselines and components, but AAP adopts learning-progress style objectives because the agenda is explicitly concerned with the rate of improvement in predictive compression, not merely exposure to unexpected inputs (Pathak et al., 2017; Burda et al., 2019; Schmidhuber, 2010).

A second pillar is empowerment and information-theoretic agency. Empowerment formalizes agent-centric control as a channel-capacity-like quantity from action sequences to future observations and provides a principled bridge between control, exploration, and interface design (Klyubin et al., 2005a,b; Salge et al., 2014; Jung et al., 2011). AAP extends this perspective by explicitly pairing empowerment with a complementary observation-to-action quantity (plasticity, proxied here by directed information) and by embedding both within a resource-cost objective. This pairing is useful because the same behavior can look desirable or undesirable depending on whether frequent reactive updates are cheap or expensive (Massey, 1990; Kramer, 1998).

AAP also draws heavily on predictive-processing, active-inference, and world-model traditions, all of which emphasize predictive internal structure and action under uncertainty (Rao and Ballard, 1999; Friston, 2010; Aitchison and Lengyel, 2017; Parr et al., 2022; Ha and Schmidhuber, 2018; Hafner et al., 2023). Most relevant is the idea that good internal models support efficient action and adaptation. AAP differs mainly in emphasis: it foregrounds explicit sensing, acting, communication, and compute bottlenecks as manipulable interface variables, and it studies how agents allocate resources across these channels rather than focusing only on the representational objective.

The agenda further overlaps with information-theoretic and thermodynamic treatments of learning and decision-making. The thermodynamics-of-prediction results of Still and colleagues motivate

the coupling between predictive information and dissipative cost (Still, 2009; Still et al., 2012), while related work on thermodynamic metrics and bounded-rational decision-making clarifies how information-processing costs can shape optimal behavior (Sivak and Crooks, 2012; Ortega and Braun, 2013). AAP imports this intuition at the level of design pressure: under viability constraints, partial observability and control, and nonzero costs for memory maintenance, action, and adaptation, prediction and selective control become economically important, not merely descriptively useful.

On the representation side, AAP is consistent with MDL and information-bottleneck perspectives, which treat useful representations as compressed summaries that preserve task- or prediction-relevant structure (Rissanen, 1978; Grünwald, 2007; Tishby et al., 1999; Cover and Thomas, 2006; Hutter, 2005; Bengio et al., 2013). Our contribution is a synthesis: predictive compression is evaluated together with control, empowerment, and interface quality under explicit budgets. This matters because predictive compression alone does not determine whether a system can alter the variables it predicts, and control alone does not ensure that the controlled variables are informative or broadly useful.

Our discussion of language and self-communication intersects several strands of work: extended cognition and cognitive offloading (Clark and Chalmers, 1998); resource-rational cognition and metareasoning (Lieder and Griffiths, 2020; Callaway et al., 2018); adaptive computation mechanisms that decide when to spend additional internal computation (Graves, 2016; Banino et al., 2021); and recent reasoning systems that partially decouple internal deliberation from public verbalization (Yao et al., 2023; Kim et al., 2025; Xie et al., 2025; Wang et al., 2026). Our position is perhaps narrower and more operational: language-like traces are one optional channel in a broader token/action budget, and their value should be measured relative to latent alternatives under matched compute and communication cost.

Finally, AAP is related to benchmark and systems-level discussions of intelligence that emphasize priors, generalization, and real-world constraints. ARC-AGI-style framing highlights the importance of priors and skill-acquisition efficiency (Chollet, 2019); broader arguments such as "Reward is Enough" emphasize the generality of reward-driven optimization mechanisms (Silver et al., 2021); and discussions of autonomous machine intelligence and human-like learning emphasize the need for richer world models and action-grounded training regimes (Lake et al., 2017; LeCun, 2022).

7 Limitations

We highlight several structural limitations and invite the community for feedback and suggestions for improvement. First, non-equivalence between prediction, empowerment, and task reward is common rather than exceptional; the agenda explicitly expects regimes in which these quantities diverge because of reward misspecification, hidden confounders, or severe partial observability. Second, some of the most interesting quantities in AAP—especially empowerment and directed information in high-dimensional settings—are expensive to estimate and may require loose variational bounds, which can blur interpretation.

A third limitation concerns measurement of energy and efficiency. FLOPs, wall-clock time, and token count are convenient but imperfect proxies for physical energy expenditure, and cross-system comparisons are sensitive to hardware, implementation details, and parallelism (Patterson et al., 2021). A fourth limitation is that self-communication channels can become brittle or degenerate into verbose private codes unless regularized by explicit bandwidth costs, latency constraints, or downstream utility. Finally, human-likeness remains intrinsically multi-objective: similarity in

constraints may improve interpretability in some settings, but it does not automatically imply safety or desirable behavior across tasks. Our manifold view of intelligence highlights the need for better fundamental understanding and science behind how specific human-related constraints and frames of reference shape and define human intelligence and how the broader manifold looks like.

8 Conclusion

We make the argument for treating AI systems as *embedded, resource-bounded agents* whose usefulness is best understood at the level of the coupled human–tool–environment system. We formalized this stance with (i) curiosity-as-learning-progress as an intrinsic development signal, (ii) explicit budgets over observation, actuation, compute, and memory in a single objective (equation (12)), (iii) complementary proxies for prediction, empowerment/plasticity, and (iv) *unification* as an experimentally manipulable notion of interface quality (equation (16)), including language and self-communication as selective information bottlenecks rather than universally privileged computation.

AAP is a call to make the tradeoffs that already govern real deployment *explicit and measurable*: what to observe, when to think, what to communicate, and how much energy and latency are worth spending for marginal gains in prediction and control. We hope this framing helps shift discussion from single-score capability to a more grounded science of agency under constraints, and we invite criticism, counterexamples, and sharper theoretical framing and experimental designs.

References

David Abel, Michael Bowling, André Barreto, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, et al. Plasticity as the mirror of empowerment. *arXiv preprint arXiv:2505.10361*, 2025.

Laurence Aitchison and Máté Lengyel. With or without you: Predictive coding and bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227, 2017. doi: 10.1016/j.conb.2017.08.010. URL `https://doi.org/10.1016/j.conb.2017.08.010`.

Andrea Banino, Jan Balaguer, and Charles Blundell. Pondernet: Learning to ponder. *arXiv preprint arXiv:2107.05407*, 2021. URL `https://arxiv.org/abs/2107.05407`.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50. URL `https://doi.org/10.1109/TPAMI.2013.50`.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *International Conference on Learning Representations*, 2019. URL `https://arxiv.org/abs/1810.12894`.

Frederick Callaway, Sayan Gul, Paul M. Krueger, Thomas L. Griffiths, and Falk Lieder. Learning to select computations. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018. URL `https://arxiv.org/abs/1711.06892`.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. URL `https://arxiv.org/abs/1911.01547`.

Andy Clark and David J. Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998. doi: 10.1093/analys/58.1.7. URL https://philpapers.org/rec/CLATEM.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X. URL https://doi.org/10.1002/047174882X.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010. doi: 10.1038/nrn2787. URL https://doi.org/10.1038/nrn2787.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016. URL https://arxiv.org/abs/1603.08983.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007. URL https://mitpress.mit.edu/9780262072816/the-minimum-description-length-principle/.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Jialin Pan, Mohammad Norouzi, Jimmy Ba, Pieter Abbeel, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, 2005. doi: 10.1007/b138233. URL https://doi.org/10.1007/b138233.

Tobias Jung, Daniel Polani, and Peter Stone. Empowerment for continuous agent–environment systems. *Adaptive Behavior*, 19(1):16–39, 2011. doi: 10.1177/1059712310392389. URL https://doi.org/10.1177/1059712310392389.

Eunki Kim, Sangryul Kim, and James Thorne. Learning to insert [pause] tokens for better reasoning. *arXiv preprint arXiv:2506.03616*, 2025. URL https://arxiv.org/abs/2506.03616.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC)*, pages 128–135, 2005a. doi: 10.1109/CEC.2005.1554676. URL https://doi.org/10.1109/CEC.2005.1554676.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. All else being equal be empowered. In *Advances in Artificial Life*, volume 3630 of *Lecture Notes in Computer Science*, pages 744–753. Springer, 2005b. doi: 10.1007/11553090_75. URL https://doi.org/10.1007/11553090_75.

Gerhard Kramer. Directed information for channels with feedback. *PhD thesis, ETH Zürich*, 1998. URL https://www.ce.cit.tum.de/fileadmin/w00cgn/lnt/staff/kramer/Papers/KramerThesis.pdf.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi: 10.1017/S0140525X16001837. URL https://doi.org/10.1017/S0140525X16001837.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview / Position Paper*, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf.

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. doi: 10.1017/S0140525X1900061X. URL https://doi.org/10.1017/S0140525X1900061X.

James L. Massey. Causality, feedback and directed information. *Proceedings of the International Symposium on Information Theory and its Applications (ISITA)*, pages 303–305, 1990. URL https://www.isiweb.ee.ethz.ch/archive/massey_pub/pdf/BI532.pdf.

Pedro A. Ortega and Daniel A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A*, 469(2153):20120683, 2013. doi: 10.1098/rspa.2012.0683. URL https://doi.org/10.1098/rspa.2012.0683.

Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2): 265–286, 2007. doi: 10.1109/TEVC.2006.890271. URL https://doi.org/10.1109/TEVC.2006.890271.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022. doi: 10.7551/mitpress/12444.001.0001. URL https://mitpress.mit.edu/9780262045353/active-inference/.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017. URL https://arxiv.org/abs/1705.05363.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. URL https://arxiv.org/abs/2104.10350.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580. URL https://doi.org/10.1038/4580.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5. URL https://doi.org/10.1016/0005-1098(78)90005-5.

Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000. doi: 10.1037/0003-066X.55.1.68. URL https://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf.

Christoph Salge, Cornelius Glackin, and Daniel Polani. Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789–2819, 2014. doi: 10.3390/e16052789. URL https://arxiv.org/abs/1406.1767.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010. doi: 10.1109/TAMD.2010.2056368. URL https://doi.org/10.1109/TAMD.2010.2056368.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. doi: 10.1016/j.artint.2021.103535. URL https://doi.org/10.1016/j.artint.2021.103535.

David A. Sivak and Gavin E. Crooks. Thermodynamic metrics and optimal paths. *Physical Review Letters*, 108(19):190602, 2012. doi: 10.1103/PhysRevLett.108.190602. URL https://doi.org/10.1103/PhysRevLett.108.190602.

Susanne Still. Information theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85 (2):28005, 2009. doi: 10.1209/0295-5075/85/28005. URL https://arxiv.org/abs/0709.1948.

Susanne Still, David A. Sivak, Anthony J. Bell, and Gavin E. Crooks. Thermodynamics of prediction. *Physical Review Letters*, 109(12):120604, 2012. doi: 10.1103/PhysRevLett.109.120604. URL https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.109.120604.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 1999. URL https://arxiv.org/abs/physics/0004057.

Jiecong Wang, Hao Peng, and Chunyang Liu. Latent chain-of-thought as planning: Decoupling reasoning from verbalization. *arXiv preprint arXiv:2601.21358*, 2026. URL https://arxiv.org/abs/2601.21358.

John Archibald Wheeler. Information, physics, quantum: The search for links. *Proceedings III International Symposium on Foundations of Quantum Mechanics*, 1989.

Zhifei Xie, Ziyang Ma, Zihang Liu, Kaiyu Pang, Hongyu Li, Jialin Zhang, Yue Liao, Deheng Ye, Chunyan Miao, and Shuicheng Yan. Mini-omni-reasoner: Token-level thinking-in-speaking in large speech models. *arXiv preprint arXiv:2508.15827*, 2025. URL https://arxiv.org/abs/2508.15827.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023. URL https://arxiv.org/abs/2305.10601.