Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Doctoral Consortium Track

A Formal Theory of Optimal Learning with Experimental Results

Michael Timothy Bennett
The Australian National University
michael.bennett@anu.edu.au

Abstract

Here I summarise some results from my thesis which are relevant to artificial intelligence (AI) and machine learning (ML). The key contribution is a theory of optimally sample and energy efficient learning, which is supported by mathematical proofs and experimental results.

1 Introduction

Some believe simplicity is the key to general intelligence [Chaitin, 1966]. Simpler models generalise more often, meaning if either of two programs could have generated data and we have to guess which, it is more likely to be the simpler of the two. This correlation between simplicity and generalisation has profound implications for AI. Yet simplicity is a property of form, not function. The function of software is determined by the hardware that interprets it. This undermines claims regarding the behaviour of the theorised, software superintelligence AIXI [Leike and Hutter, 2015]. In theory there could be no correlation between simplicity and generalisation. In practice, there is. Why?

This question cannot be answered by formalisms like AIXI [Hutter and others, 2024], because they frame AI as a disembodied policy interacting with the environment through an interpreter. In previous work I named this 'computational dualism' [Bennett, 2024a], and argued an alternative is needed. The full formal treatment is lengthy, but greatly simplified version of the argument is as follows. Assume $\mathcal{C}$ is a space of software minds like AIXI. I want the optimal $f_1 \in \mathcal{C}$. I can't assume details like actions or timesteps because that would presuppose aspects of interpretation and undermine any claim I subsequently make [Leike and Hutter, 2015]. Fortunately we only need to know if overall behaviour satisfies goals, not how. Γ is a set of overall behaviours. $f_2 : \mathcal{C} \to \Gamma$ is a hardware body, like the UTM AIXI uses. It interprets f_1. $f_3 : \Gamma \to \{0, 1\}$ is the environment in which goals are pursued. If I constructed f_1 for a purpose, then *I* decide whether it has fulfilled that purpose, and *I* am part of its environment f_3. Goals are satisfied iff $f_3(f_2(f_1)) = 1$. The performance of f_1 in f_3 depends on f_2. It is pointless to make claims based on f_1 alone. That is *computational dualism*. The problem is resolved by formalising the cognitive system-as-a-whole. I call this formalism Pancomputational Enactivism.

2 Formalism

Pancomputationalism holds all systems are computational [Piccinini, 2015]. Enactivism rejects the agent-environment paradigm, reframing cognition as a part of the environment [Thompson, 2007]. In Pancomputational Enactivism, goals are indistinguishable from the systems that pursue them. Instead, a 'task' represents both goal and computational system. Tasks allow claims to be made regarding *subjective* complexity of form in relation to *objective* function. Non-standard definitions of terms like language and extension are used to avoid computational dualism. All environments have at least one 'state', states are defined by their relative differences, and time is difference (so states are mutually exclusive). The justification for all of this is given at length in the thesis [Bennett, 2025a]. With abuse of notation, the formalism is:

- **Programs:** Φ is the set of *states*, and $P = 2^{\Phi}$ is all possible *declarative programs*. Each program is a potential point of difference between states. Assume a present state $\phi \in \Phi$, and program $f \in P$ is true if $\phi \in f$.
- **Embodied Language:** A body is formalised as a finite *vocabulary* $\mathfrak{v} \subset P$. The embodied language is $L_{\mathfrak{v}} = \{l \subseteq \mathfrak{v} : \bigcap l \neq \emptyset\}$. Members of $L_{\mathfrak{v}}$ are *statements* the body can express (state of memory etc). A statement $l \subseteq L_{\mathfrak{v}}$ is true when $\phi \in \bigcup l$. $E_l = \{y \in L_{\mathfrak{v}} : l \subseteq y\}$ is called the **extension** of l.
- **$\mathfrak{v}$-Tasks:** A $\mathfrak{v}$-task $\alpha = \langle I_\alpha, O_\alpha \rangle$ has inputs $I_\alpha \subset L_{\mathfrak{v}}$ and outputs $O_\alpha \subset E_{I_\alpha}$. Assume a uniform distribution over $\mathfrak{v}$-tasks. Let α and ω be $\mathfrak{v}$-tasks. If $I_\alpha \subset I_\omega$, $O_\alpha \subseteq O_\omega$, then α is examples of ω.
- **Policies:** $\pi \in L_{\mathfrak{v}}$ is a correct policy[1] for α if $E_{I_\alpha} \cap E_\pi = O_\alpha$. Π_α is the set of all correct policies for α. The embodied system learns or generalises to ω by inferring a correct policy from examples α that is also correct for ω, meaning $\pi \in \Pi_\alpha \cap \Pi_\omega$. 'Experience' is adding inputs and outputs to the examples α. Intelligence is measured by sample and energy efficiency in learning ω from α.
- **Heuristics:** The weakness of a policy $\pi \in \Pi_\alpha$ is $|E_\pi|$, and its complexity is the cardinality of the smallest correct policy $\pi' \in \Pi_\alpha$ s.t. $E_{\pi'} = E_\pi$[2].

[1] Assume only policies that imply past observations, like hypotheses in a Boolean SAT problem [Russell and Norvig., 2021].

[2] Basically minimum description length [Rissanen, 1978]. Intu-

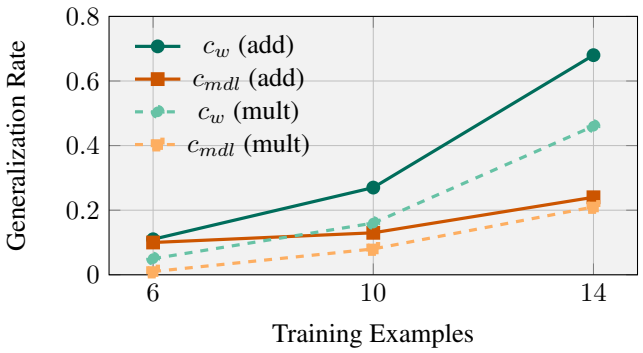

Figure 1: Rates for binary addition (solid) and mult. (dashed).

3 Results

Proof results. Assume you must choose a policy $\pi \in \Pi_\alpha$. I show that to maximise the probability of choosing $\pi \in \Pi_\alpha \cap \Pi_\omega$, maximising $|E_\pi|$ is necessary and sufficient. It is neither necessary nor sufficient to minimise complexity [Bennett, 2023b; Bennett, 2025c]. I prove an *objective* upper bound for sample and energy efficiency in adaptation, to complement AIXI's subjective upper bound [Bennett, 2024a]. I show correlation between simplicity and generalisation can be explained by weakness confounding the two [Bennett, 2024b]. These results are summarised as an epistemological razor: "explanations should be no more specific than necessary". See thesis appendices for proofs [Bennett, 2025a].

Experimental results. I implemented an A* search based learning system and had it learn to predict binary strings (4-bits input, 4-bits output) [Bennett, 2023b]. The tasks were binary multiplication and addition. I compared weakness and simplicity as heuristics for policy selection. Experiments were run with different numbers of examples to learn from. In each experiment search found a simplest (c_{mdl}) and a weakest (c_w) policy to compare. With 6–14 training examples (out of 16 total), choosing weak policies yielded a $110 - 500\%$ improvement in generalisation rate (Figure 1). See GitHub for details [Bennett, 2025a].

Interpretation. This can be seen as an alternative to the minimum description length principle [Rissanen, 1978]. The proofs and experiments show weakness is the better heuristic. This shows how to build more energy and sample efficient AI. If scaled, this approach will yield more reliable systems useful for domains where error-tolerance is low. Due to the enactive frame, this formalism also addresses questions in philosophy, biology and complex systems. Two significant results have already been obtained. The first is a formal explanation of the evolution and function of consciousness [Bennett, 2023a; Bennett *et al.*, 2024]. The second formalises the multilayer architecture of biological self-organisation to show these systems leverage the delegation of control to maximise adaptability [Bennett, 2025c]. This provides a formal foundation for 'soft robotics' [Man and Damasio, 2019] and neuromophic computing [Borghi *et al.*, 2024]. These results all together form a thesis [Bennett, 2025a] at the intersection of AI, philosophy of mind and theoretical neurobiology in pursuit of general intelligence [Goertzel, 2014; Bennett, 2025b].

itively, weakness measures function and complexity measures form.

References

[Bennett *et al.*, 2024] Michael T. Bennett, Sean Welsh, and Anna Ciaunica. Why is anything conscious? *Preprint*, 2024. https://osf.io/preprints/osf/mtgn7_v2.

[Bennett, 2023a] Michael T. Bennett. Emergent causality and the foundation of consciousness. In *Artificial General Intelligence*. Springer Nature, 2023.

[Bennett, 2023b] Michael T. Bennett. The optimal choice of hypothesis is the weakest, not the shortest. In *Artificial General Intelligence*. Springer Nature, 2023.

[Bennett, 2024a] Michael T. Bennett. Computational dualism and objective superintelligence. In *Artificial General Intelligence*. Springer, 2024.

[Bennett, 2024b] Michael T. Bennett. Is complexity an illusion? In *Artificial General Intelligence*. Springer, 2024.

[Bennett, 2025a] Michael T. Bennett. *How To Build Conscious Machines*. PhD thesis, School of Computing, The Australian National University, 2025. github.com/ViscousLemming/Technical-Appendices.

[Bennett, 2025b] Michael T. Bennett. What the f*ck is artificial general intelligence? In *Artificial General Intelligence (forthcoming)*. Springer Nature, 2025.

[Bennett, 2025c] Michael Timothy Bennett. Are biological systems more intelligent than artificial intelligence? 2025. Forthcoming.

[Borghi *et al.*, 2024] Francesca Borghi, Thierry R. Nieus, Davide E. Galli, and Paolo Milani. Brain-like hardware, do we need it? *Frontiers in Neuroscience*, 18, 2024.

[Chaitin, 1966] Gregory J. Chaitin. On the length of programs for computing finite binary sequences. *J. ACM*, 1966.

[Goertzel, 2014] Ben Goertzel. Artificial general intelligence: Concept, state of the art. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014.

[Hutter and others, 2024] Marcus Hutter et al. *An Introduction to Universal Artificial Intelligence*. CRC, 2024.

[Leike and Hutter, 2015] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. *Proceedings of The 28th Conference on Learning Theory*, 2015.

[Man and Damasio, 2019] Kingson Man and Antonio R. Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 2019.

[Piccinini, 2015] Gualtiero Piccinini. *Physical Computation: A Mechanistic Account*. Oxford Press, 2015.

[Rissanen, 1978] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[Russell and Norvig., 2021] S. Russell and P. Norvig. *Artificial intelligence: A modern approach, 4th ed.* 2021.

[Thompson, 2007] Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Havard, 2007.