Enactivism & Objectively Optimal Super-Intelligence

Michael Timothy Bennett[1][0000-0001-6895-8782]

The Australian National University michael.bennett@anu.edu.au

Abstract. Software's effect upon the world hinges upon the hardware that interprets it. This tends not to be an issue, because we standardise hardware. AI is typically conceived of as a software "mind" running on such interchangeable hardware. This formalises mind-body dualism, in that a software "mind" can be run on any number of standardised bodies. While this works well for simple applications, we argue that this approach is less than ideal for the purposes of formalising artificial general intelligence (AGI) or artificial super-intelligence (ASI). The general reinforcement learning agent AIXI is pareto optimal. However, this claim regarding AIXI's performance is highly subjective, because that performance depends upon the choice of interpreter. We examine this problem and formulate an approach based upon enactive cognition and pancomputationalism to address the issue. Weakness is a measure of plausibility, a "proxy for intelligence" unrelated to compression or simplicity. If hypotheses are evaluated in terms of weakness rather than length, then we are able to make objective claims regarding performance (how effectively one adapts, or "generalises" from limited information). Subsequently, we propose a definition of AGI which is objectively optimal given a "vocabulary" (body etc) in which cognition is enacted, and of ASI as that which finds the optimal vocabulary for a purpose and then constructs an AGI[1].

1 Introduction

AIXI [2] provides us with a mathematically precise notion of AGI. Its performance is measured according to Legg-Hutter intelligence [3], a proxy for "the ability to satisfy goals in a wide range of environments" [4]. It employs Solomonoff Induction [5, 6] to make accurate inferences from minimal data. Because of this it is pareto optimal, meaning there is no agent which outperforms AIXI in one environment and equals its performance in all others. Unfortunately, this claim is highly subjective, because it depends upon the choice of Universal Turing Machine (UTM) [7]. We explore this problem, and formulate an approach that combines enactive cognition [8], pancomputationalism [9] and weakness as a proxy for intelligence [10, 1].

[1] Technical appendices are available on GitHub [1].

1.1 An informal explanation of AIXI

Our purpose is to explain the aforementioned subjectivity and how it might be addressed, rather than every detail of how AIXI functions. This paper is as philosophical as it is mathematical. As such, the following explanation of AIXI is informal and involves significant abuse of notation.

Models: A model can be understood as a program [2, 11] or set of rules [10] describing how aspects of the world relate to one another. A model can be used as a hypothesis, to *explain* aspects of the present by pointing out which aspects of the past caused the present [12]. Likewise, the more distant past can explain the more recent past, and the present can explain the future. Of course, a model of the world is not the world itself. Some models will more accurately represent the world than others. AIXI is not a comment on desirable behaviour, values or goals, but is built upon the assumption that such things are measured by a reward function that is given [13]. To satisfy goals, AIXI must predict[2] the consequences of its actions. To make predictions, an agent requires a model. If a model approximates the environment well enough, then the agent can accurately predict the consequences of its actions, and so form a plan that will cause its goals to become satisfied. The more accurate a model is, the more likely an agent will be able to satisfy its goals. AIXI is able to satisfy goals because it has a means of discerning which models will be most accurate.

Universal priors: How AIXI obtains an accurate representation of the world can be informally understood in two parts[3]. First, AIXI considers only models that explain the past and present precisely (by which we mean that each model is a lossless archive of past and present). Any model that would predict a different outcome to past events than what actually took place is discarded, leaving AIXI only with models consistent with what it knows to be true. While these models are equivalent with respect to the past, they may differ in what future they predict. AIXI must identify which of those models most accurately predicts the future. For this purpose it is assumed that simpler models are more plausible representations of the world (in line with Ockham's Razor [14]). Simplicity is measured in terms of Kolmogorov Complexity (KC) [15]. The KC of an object is the length of the shortest self extracting archive of that object. To give some intuition as to what this means, there may exist many models that behave in exactly the same manner in all circumstances. Those models are really the same model represented in different ways, and KC is the length of its shortest representation in a language. Models with smaller KC tend to make more accurate predictions, formalising Ockham's Razor. This is why some believe that compression and intelligence are closely related [16], because compression can be

[2] To accurately predict the future means to infer which future among possible futures has the highest probability of occurring.

[3] Again, we must emphasise that this explanation is informal - the point is just to provide some context to explain the problem of subjectivity.

used to measure simplicity and so identify explanations that are more likely to be true. AIXI prefers models that have smaller KC, and in doing so maximises the accuracy of its predictions[4]. AIXI estimates one thing (model accuracy), by measuring another seemingly unrelated thing (KC). In other words, it uses compression as a proxy. This proxy for intelligence (defined in terms of the ability to satisfy goals across a wide range of environments) gives AIXI what is called "a universal prior" [5, 6], a means of deciding which among valid models are best. This is also why AIXI is also called a *universal* artificial intelligence [17]. So to reiterate, AIXI's intelligent behaviour stems from an accurate model. How AIXI obtains an accurate model can be understood (very informally) in two steps:

1. Collect models whose predictions are consistent with what we've observed[5].
2. Use a proxy for intelligence (Kolmogorov Complexity) to decide which among those models will most accurately predict the future.

1.2 Subjectivity

KC is measured in the context of a UTM [7]. By itself, changing the UTM would not meaningfully affect performance. When used in a universal prior to predict deterministic binary sequences, the number of incorrect predictions a model will make is bounded by a multiple of the KC of that model [18]. If the UTM is changed the number of errors only changes by a constant [19, pp. 2.1.1 & 3.1.1], so changing the UTM doesn't change which model is considered most plausible. However, when AIXI employs this prior in an *interactive* setting, a problem occurs [7]. To explain in simplified terms (with abuse of notation), assume a program f_1 is software, f_2 is an interpreter and f_3 is the reality (an environment, body etc) within which goals are pursued. AIXI is the optimal choice of f_1 to maximise the performance of $f_3(f_2(f_1))$. However, in an interactive setting one's perception of success may not match reality.

> "Legg-Hutter intelligence [3] is measured with respect to a fixed UTM. AIXI is the most intelligent policy if it uses the same UTM." [7, p.10]

If intelligence is measured with respect to one UTM while AIXI runs on another, then this is like AIXI being engaged in one reality, while success is determined by another, entirely different reality. Using our analogy of functions, performance in terms of $f_3(f_2(f_1))$ depends upon $f_2(f_1)$, not f_1 alone. Thus a claim regarding the performance of f_1 alone would be *subjective*, in that it depends upon f_2.

> "This undermines all existing optimality properties for AIXI." [7, p.1]

[4] This is a simplification. More formally, if the model which generated past data is indeed computable, then the simplest model will dominate the Bayesian posterior as more and more data is observed. Eventually, you will have identified the correct model and can use that model to generate the next sample (predict the future).

[5] Meaning they all "predict" the exact same past.

A UTM is an interpreter. As Leike and Hutter pointed out, Legg-Hutter intelligence is measured with respect to a fixed interpreter. The problem disappears if AIXI uses that same interpreter, which is easier said than done. This paper explores how we might formalise cognition in a different manner, so that performance is indepenedent of the choice of interpreter. To do so we need to formalise the mind as part of the environment, and the environment as software. Using the analogy from earlier, this would give us $f_2(f_3(f_1))$ instead of $f_3(f_2(f_1))$. In that case, performance would then be measured in terms of $f_3(f_1)$, and would be unaffected by interpreter f_2.

2 Formalising Enactivism

AI is typically conceived of as a software "mind" running on an interchangeable hardware body. The hardware interacts with an environment, and the software interacts with the hardware. This formalises mind-body dualism, in that we could take the software "mind" and run it on any number of different bodies. However, this portrayal of cognition is flawed. What computer code does depends on the interpreter, we just tend to standardise system architectures. An alternative to dualism is enactivism [8] which holds that mind and body are inseparable, embedded in time and place. Cognitive activity extends into the environment, and is enacted through what the organism does. For example, if someone uses pen and paper to compute the solution to a math problem, then their cognition is extending into and enacted within the environment [20]. Formalising enactivism can address problems associated with dualism. However it is unclear how enactive cognition might work computationally, because it blurs the boundary between the agent and environment. To address this, we look to pancomputationalism [9]. Pancomputationalism holds that everything is a computational system. It follows that we may regard the interpreter f_2 as the universe, and the environment f_3 as software that runs on f_2. Consequently we have $f_2(f_3...)$ rather than $f_3(f_2...)$. The distinction between mental (software) and physical (hardware) can be discarded. This means we need to represent the model f_1 as a part of the environment f_3. We do so by merging agent, body and environment into a task [10], formalising instances of intent in such a way as may bear resemblance to Heidegger's Dasein (Being-in-the-world and bound by context) [21].

2.1 A model of the environment within the environment

There exists an isomorphism between declarative and imperative programs (the Curry-Howard isomorphism [22]). As such, we may treat both the model f_1 and then environment f_3 as declarative programs. Assume a set of declarative programs represents the logical conjunction of its members. Then, for every set of declarative programs there exists a declarative program which is equivalent. If f_1 and f_3 are sets, we can define f_1 as a subset of f_3 to represent the model as part of the environment. Because $f_1 \subset f_3$, the ability to satisfy goals is now measured in terms of $f_2(f_3)$, we can now reason about the model in objective terms. Going forward we'll discard f_3, f_2 and f_1 in favour of more formal notation, and will refer to the UTM f_2 as the pancomputationalist's universe.

Definition 1 (environment).

- *We assume a set Φ whose elements we call* ***states****, one of which we single out as the* ***present state***[6].
- *A* ***declarative program*** *is a function $f : \Phi \rightarrow \{true, false\}$, and we write P for the set of all declarative programs. By an* ***objective truth*** *about a state ϕ, we mean a declarative program f such that $f(\phi) = true$.*

2.2 We need only model the task, not all of the environment

Enactivism blurs the line between agent and environment, making the distinction unclear. As such, we abandon these separate notions entirely. The distinction is a convenient but unnecessary abstraction [10]. As Heidegger maintained, Being is bound by context [21]. There is no need to define an agent that has no environment, and so there seems to be little point in preserving the distinction. Furthermore, we do not need a model of the environment.

"The best model of the world is the world itself." - Rodney Brooks [23]

The only aspects of the environment that we might actually need model are those necessary to satisfy goals [24]. What is needed is not a model of the environment but a model describing how to satisfy a goal *while* embodied and embedded in a particular local environment. Rather than the environment, we model a task. Intuitively, a task might be seen as the instantiation of intent. To avoid confusion going forward we will refer to "the mechanism" by which decisions are made, instead of "the agent". Where a model of an environment may include details needed to predict the environment but not satisfy goals, a model of a task can ignore anything which is not necessary to satisfy the goal. As a result, a separate description of a goal is unnecessary because it is implied by which aspects of the environment are modelled. If we only need to model those aspects of the environment necessary to complete a task, then we are dealing with the necessarily finite physical circuitry with which cognition is enacted. We can represent that circuitry using a finite subset of P (the set of all declarative programs as per definition 1). This finite circuitry is a language, albeit one whose meanings are implemented in the pancomputationalist's universe rather than interpreted by a human mind. It will be used to formally describe tasks.

Definition 2 (implementable language).

- *$\mathfrak{V} = \{V \subset P : V \text{ is } finite\}$ is a set whose elements we call* ***vocabularies****, one of which*[7] *we single out as* ***the vocabulary*** *$\mathfrak{v}$.*

[6] Each state is just reality from the perspective of a point along one or more dimensions. States of reality must be separated by something, or there would be only one state. E.G. two different states may be reality at two different points in time.

[7] The vocabulary $\mathfrak{v}$ we single out represents the sensorimotor circuitry with which an organism enacts cognition - their brain, body, local environment and so forth. It is finite because the time, memory, precision etc available are assumed to be finite.

- $L_\mathfrak{v} = \{l \subseteq \mathfrak{v} : \exists \phi \in \Phi \; (\forall p \in l : p(\phi) = true)\}$ *is a set whose elements we call* ***statements****.* $L_\mathfrak{v}$ *follows* Φ *and* $\mathfrak{v}$*, and is called* ***implementable language****.*
- $l \in L_\mathfrak{v}$ *is* ***true*** *iff the present state is* ϕ *and* $\forall p \in l : p(\phi) = true$*.*
- *The* ***extension of a statement*** $a \in L_\mathfrak{v}$ *is* $Z_a = \{b \in L_\mathfrak{v} : a \subseteq b\}$*.*
- *The* ***extension of a set of statements*** $A \subseteq L_\mathfrak{v}$ *is* $Z_A = \bigcup_{a \in A} Z_a$*.*

(Notation) *Z with a subscript is the extension of the subscript*[8]*.*

The programs in $\mathfrak{v}$ are the circuitry with which cognition is enacted, and only programs in $\mathfrak{v}$ affect decision making. We assume cognition always takes place in the context of a physical machine or sensorimotor system, represented by the implementable language. With these, we can define a task.

Definition 3 ($\mathfrak{v}$-task). *For a chosen* $\mathfrak{v}$*, a task*[9] α *is* $\langle S_\alpha, D_\alpha, M_\alpha \rangle$ *where:*

- $S_\alpha \subset L_\mathfrak{v}$ *is a set whose elements we call* ***situations*** *of* α*.*
- S_α *has the extension* Z_{S_α}*, whose elements we call* ***decisions*** *of* α*.*
- $D_\alpha = \{z \in Z_{S_\alpha} : z \text{ is correct}\}$ *is the set of all decisions which complete* α*.*
- $M_\alpha = \{l \in L_\mathfrak{v} : Z_{S_\alpha} \cap Z_l = D_\alpha\}$ *whose elements we call* ***models*** *of* α*.*

$\Gamma_\mathfrak{v}$ *is the set of all tasks for our chosen* $\mathfrak{v} \in \mathfrak{V}$*.*

(Notation) *If* $\omega \in \Gamma_\mathfrak{v}$*, then we will use subscript* ω *to signify parts of* ω*, meaning one should assume* $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ *even if that isn't written.*

(How a task is completed) *Assume we've a* $\mathfrak{v}$*-task* ω *and a hypothesis* $\boldsymbol{h} \in L_\mathfrak{v}$ *s.t.*

1. *we are presented with a situation* $s \in S_\omega$*, and*
2. *we must select a decision* $z \in Z_s \cap Z_{\boldsymbol{h}}$*.*
3. *If* $z \in D_\omega$*, then* z *is correct and the task is complete. This occurs if* $\boldsymbol{h} \in M_\omega$*.*

$\omega \in \Gamma_\mathfrak{v}$ s.t. $S_\omega \subset S_\alpha$, $D_\omega \subset D_\alpha$ and $D_\omega \subset Z_{S_\omega}$ can serve as an ostensive definition [25] of α from which to infer $\mathbf{h}$. Then, if $\mathbf{h} \in M_\alpha$, then $z \in D_\alpha$.

A solitary decision instead of a sequence: Where AIXI deals in sequential decisions over time [2], a $\mathfrak{v}$-task is completed with only one. This is because:

1. For every sequence of decisions there exists an equivalent single decision, in much the same way as any planning problem can be represented as a boolean satisfiability problem [26]. Not all tasks involve a sequence, but all involve at least one decision. If a single decision will suffice, why complicate matters?
2. A single decision may set in motion continuous interactions. The preference for sequences may suit reinforcement learners using discrete, pre-defined actions, however in the enactive context such abstractions are not given.
3. Whether behaviour is the result of one decision or many does not matter. What matters is whether the task is completed as a result.

[8] E.G. Z_s is the extension of s.

[9] E.G. this could represent chess as a supervised learning problem where $s \in S_\alpha$ is the state of a chessboard, $z \in Z_s$ is a sequence of moves by two players that begins in s, and $d \in D_\alpha \cap Z_s$ is such a sequence of moves that resulted in victory for one player in particular (the one undertaking the task).

Binary correctness: To further simplify matters, correctness is binary. Given a task, a decision is considered to be either correct or incorrect. It may be that a decision is correct if it causes the task to become complete to some acceptable degree with some acceptable probability – what is otherwise known as satisficing [27][10]. Degrees of complete or correct just reflect different task definitions. Preferences that determine what is considered complete, methods of attributing task completion to past decisions are beyond this paper's scope. Preferences and the emergence of identity are formalised in a companion to this paper [28, 29].

Representing the past to predict the future: Earlier we described (very informally) how an accurate model can be obtained by discarding any model that "predicts" a different outcome from past events than what eventuated, and then using a proxy for intelligence to determine which among those that remain will most accurately predict the future. There exists a set of all decisions that *might* ever be made which are correct, which we can use to specify a $\mathfrak{v}$-task ω. Likewise, the past can be represented as the set of all decisions that *have* been made. From the past we can construct an ostensive definition of ω, by specifying a $\mathfrak{v}$-task α where D_α is the set of all decisions which *have* been made *and* were deemed correct, given the situations S_α in which they were made (this assumes a means of attributing correctness to past decisions). For each $m \in M_\alpha$ the past is $Z_m \cap Z_{S_\alpha} = D_\alpha$ and the future decisions implied are $Z_m \cap (Z_{S_\omega} - Z_{S_\alpha})$ (the decisions implied for all situations that have not yet been experienced). In other words, the models in M_α are equivalent with respect to the past but may disagree about the future. We know $D_\alpha \subset D_\omega$, so the larger $|Z_m \cap D_\omega|$ is, the more accurate m's predictions. We would use a proxy for intelligence to determine which $m \in M_\alpha$ is most accurate.

3 The objectively optimal hypothesis

Having formulated cognition as a task, merging agent and environment, we have ensured that any claims regarding performance are now unaffected by the choice of interpreter. This addresses subjectivity as it pertained to AIXI. Unfortunately, it introduces other problems we must now address. First, Legg-Hutter intelligence is not well defined for a task. Second, we can no longer use Kolmogorov Complexity because everything must be represented in the same implementable language. We could use minimum description length [30] (compressing data written using a vocabulary to an archive written using that same vocabulary), however selecting hypotheses by length would still render claims about performance subjective. Third, we must now show that not only is an optimal hypothesis objectively so given $\mathfrak{v}$, but define the objectively optimal choice of $\mathfrak{v}$. For this we

[10] Alternatively, feedback (feelings / reward signals) may be given through the implementable language (as declarative programs), and each situation expresses a threshold with respect to what is considered "good enough".

require a measure of performance, and an alternative proxy for intelligence. Both are addressed [31] by companion to this paper concerning optimal hypotheses.

3.1 Performance

AIXI is asymptotically optimal, meaning given enough data on the past it will predict the future accurately (adapting to its environment). However, because only finitely many tasks can be expressed in an implementable language that yardstick is no longer particularly meaningful (we can rote learn a finite set). Instead, we will define performance in terms of how *quickly* a mechanism adapts. Thus we take performance to be the ability to generalise from limited information[11], citing arguments to the effect that such a thing is intelligence [10, 11].

Definition 4 (generalisation). *A statement l generalises to $\alpha \in \Gamma_{\mathfrak{v}}$ iff $l \in M_\alpha$, because then $D_\alpha = Z_l \cap S_\alpha$. We say l generalises from α to $\mathfrak{v}$-task ω if we first obtain l from M_α and then find it generalises to ω.*

We assume a uniform distribution over $\Gamma_{\mathfrak{v}}$. The probability that $l \in L_{\mathfrak{v}}$ generalises to a randomly sampled $\mathfrak{v}$-task ω is $p(l \in M_\omega \mid l \in L_{\mathfrak{v}}) = \frac{2^{|Z_l|}}{2^{|L_{\mathfrak{v}}|}}$. Assume α and ω are $\mathfrak{v}$-tasks s.t. $S_\alpha \subset S_\omega$, $D_\alpha \subset D_\omega$ and $D_\alpha \subset Z_{S_\alpha}$. We wish to generalise from α to ω[12]. The mechanism selects a hypothesis $\mathbf{h} \in M_\alpha$, and performance is measured as $p(\mathbf{h} \in M_\omega \mid \mathbf{h} \in M_\alpha)$.

3.2 Weakness as a proxy for intelligence

First, we must explain why description length is an unsuitable proxy. In the context of $m \in M_\alpha$ description length [30] might be most faithfully translated as the cardinality $|m|$ of m. For every conceivable task α there exists a program $u \in P$ such that $Z_{\{u\}} = D_\alpha$. If $u \in \mathfrak{v}$ then the minimum description length model is $\{u\}$ and $p(m \in M_\omega \mid m \in M_\alpha) = 0$. Hence, minimising description length does not guarantee optimal performance. Any claim regarding the performance of a mechanism using length as a proxy would still be subjective. Instead of $|m|$ we can use $|Z_m|$ (the cardinality of m's extension Z_m), called the "weakness" of m. It is arguable that intelligence is a measure of the ability to generalise from one task to another, which amounts to a preference to weaker hypotheses [10]. If tasks are uniformly distributed then the probability of a statement l generalising to an unknown task ω proportional to l's weakness. If we use weakness as our proxy (to choose between models) instead of description length, then optimal performance is attained by choosing $\mathbf{h} \in \underset{m \in M_\alpha}{\arg\max}\ p(m \in M_\omega \mid m \in M_\alpha) = \underset{m \in M_\alpha}{\arg\max}\ |Z_m|$. There is no choice of $\mathfrak{v}$ which can make weaker models less likely to generalise, because one cannot increase $|Z_m|$ without increasing $\frac{2^{|Z_m|}}{2^{|L|}}$. In contrast $|m|$ need not bear any relationship to $\frac{2^{|Z_m|}}{2^{|L|}}$. It follows that $\mathbf{h} \in \underset{m \in M_\alpha}{\arg\max}\ |Z_m|$ is objectively optimal, in the sense that it is optimal given any choice of either $\mathfrak{v}$ or ω.

[11] The speed of adaptation, or how few examples one needs to understand a concept.
[12] In the absence of knowledge α, $p(\mathbf{h} \in M_\omega \mid \mathbf{h} \in L_{\mathfrak{v}})$ is maximised when $l = \emptyset$.

3.3 Objectively optimal AGI and ASI

Given the above and related arguments [17, 32], we propose defining AGI and ASI as follows. These are mathematical ideals we may aim to build, rather than an approach to doing so. An AGI is an agent that selects the optimal hypothesis for any given task. An ASI selects the optimal vocabulary to maximise the utility of intelligence for a task, and then implements an AGI with that vocabulary. If $\mathbf{h} \in L_{\mathfrak{v}}$ is our AGI's hypothesis and $\alpha \in \Gamma_{\mathfrak{v}}$ its knowledge[13], then $\mathbf{h} \in \underset{m \in M_\alpha}{\arg\max}\, |Z_m|$. Let $\Gamma_{\mathfrak{V}} = \bigcup_{\mathfrak{k} \in \mathfrak{V}} \Gamma_{\mathfrak{k}}$ be the set of all tasks across all vocabularies. Let λ be a function $\lambda : \mathfrak{V} \rightarrow \Gamma_{\mathfrak{V}}$ that takes a vocabulary and returns a task in that vocabulary. λ lets us represent a version of the same task in different vocabularies. Every task $\gamma \in \Gamma_{\mathfrak{V}}$ has a utility of intelligence value (how useful it is), computed by $\epsilon : \Gamma_{\mathfrak{V}} \rightarrow \mathbb{N}$ s.t. $\epsilon(\gamma) = \underset{m \in M_\gamma}{\arg\max}(|Z_m| - |D_\gamma|)$. If λ is our ASI's knowledge and $\mathbf{h}$ its hypothesis, then it uses $\mathfrak{v}$ s.t.

$$\mathfrak{v} \in \underset{\mathfrak{v} \in \mathfrak{V}}{\arg\max}\ \epsilon\left(\lambda(\mathfrak{v})\right) \text{ and } \mathbf{h} \in \underset{m \in M_{\lambda(\mathfrak{v})}}{\arg\max}\ |Z_m|$$

If $\mathfrak{J} \subseteq \mathfrak{V}$ is the set of vocabularies for which ϵ has been computed, then an anytime computable alternative is $\mathfrak{v} \in \underset{\mathfrak{v} \in \mathfrak{J}}{\arg\max}\ \epsilon(\lambda(\mathfrak{v}))$.

References

[1] M. T. Bennett. *Technical Appendices*. Version 1.2.1. 2023. DOI: `10.5281/zenodo.7641742`. URL: `https://github.com/ViscousLemming/Technical-Appendices` (visited on 03/04/2023).

[2] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin, Heidelberg: Springer-Verlag, 2010.

[3] S. Legg and M. Hutter. "Universal Intelligence: A Definition of Machine Intelligence". In: *Minds and Machines* 17.4 (2007), pp. 391–444.

[4] S. Legg. "Machine Super Intelligence". PhD thesis. Uni. of Lugano, 2008.

[5] R. Solomonoff. "A formal theory of inductive inference. Part I". In: *Information and Control* 7.1 (1964), pp. 1–22.

[6] R. Solomonoff. "A formal theory of inductive inference. Part II". In: *Information and Control* 7.2 (1964), pp. 224–254.

[7] J. Leike and M. Hutter. "Bad Universal Priors and Notions of Optimality". In: *Proceedings of The 28th Conference on Learning Theory, in Proceedings of Machine Learning Research* (2015), pp. 1244–1259.

[8] D. Ward, D. Silverman, and M. Villalobos. "Introduction: The Varieties of Enactivism". In: *Topoi* 36 (Apr. 2017).

[9] G. Piccinini and C. Maley. "Computation in Physical Systems". In: *The Stanford Encyclopedia of Philosophy*. Sum. 21. Stanford University, 2021.

[13] This assumes either that knowledge only consists of an ostensive definition of what "good enough" is, or that feedback is programs in $\mathfrak{v}$, and that each situation expresses a threshold with respect to what is considered "good enough".

[10] M. T. Bennett. “Symbol Emergence and the Solutions to Any Task”. In: *Artificial General Intelligence*. 2022, pp. 30–40.
[11] F. Chollet. *On the Measure of Intelligence*. 2019.
[12] M. T. Bennett. “Computable Artificial General Intelligence”. In: *Manuscript* (May 2022).
[13] M. T. Bennett and Y. Maruyama. “Philosophical Specification of Empathetic Ethical Artificial Intelligence”. In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2022), pp. 292–300.
[14] E. Sober. *Ockham’s Razors: A User’s Manual*. Cambridge Uni. Press, 2015.
[15] A. Kolmogorov. “On tables of random numbers”. In: *Sankhya: The Indian Journal of Statistics* A (1963), pp. 369–376.
[16] G. Chaitin. “The Limits of Reason”. In: *Sci. Am.* 294.3 (2006), pp. 74–81.
[17] B. Goertzel. “Artificial General Intelligence: Concept, State of the Art”. In: *Journal of Artificial General Intelligence* 5.1 (2014), pp. 1–48.
[18] R. Solomonoff. “Complexity-based induction systems: Comparisons and convergence theorems”. In: *IEEE Transactions on Information Theory* 24.4 (1978), pp. 422–432.
[19] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications (Third Edition)*. New York: Springer, 2008.
[20] A. Clark and D. J. Chalmers. “The Extended Mind”. In: *Analysis* 58.1 (1998), pp. 7–19.
[21] M. Wheeler. “Martin Heidegger”. In: *The Stanford Encyclopedia of Philosophy*. Fall 2020. Stanford University, 2020.
[22] W. A. Howard. “The Formulae-as-Types Notion of Construction”. In: *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*. Academic Press, 1980, pp. 479–490.
[23] H. L. Dreyfus. “Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian”. In: *Philosophical Psychology* 20.2 (2007), pp. 247–268.
[24] M. T. Bennett. “Compression, The Fermi Paradox and Artificial Super-Intelligence”. In: *Artificial General Intelligence*. 2022, pp. 41–44.
[25] A. Gupta. “Definitions”. In: *The Stanford Encyclopedia of Philosophy*. Winter 2021. Stanford University, 2021.
[26] H. Kautz and B. Selman. “Planning as satisfiability”. In: *IN ECAI-92*. New York: Wiley, 1992, pp. 359–363.
[27] F. M. Artinger, G. Gigerenzer, and P. Jacobs. “Satisficing: Integrating Two Traditions”. In: *Journal of Economic Literature* 60.2 (2022), pp. 598–635.
[28] M. T. Bennett. *Emergent Causality & the Foundation of Consciousness*. 2023. URL: https://arxiv.org/abs/2302.03189.
[29] M. T. Bennett. *On the Computation of Meaning, Language Models and Incomprehensible Horrors*. 2023.
[30] J. Rissanen. “Modeling By Shortest Data Description*”. In: *Autom.* 14 (1978), pp. 465–471.
[31] M. T. Bennett. *The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest*. 2023. URL: https://arxiv.org/abs/2301.12987.

[32] M. T. Bennett and Y. Maruyama. "The Artificial Scientist: Logicist, Emergentist, and Universalist Approaches to Artificial General Intelligence". In: *Artificial General Intelligence*. 2022, pp. 45–54.