Enactivism & Objectively Optimal Super-Intelligence

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

The Australian National University michael.bennett@anu.edu.au
http://www.michaeltimothybennett.com/

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. The hardware interacts with an environment, and the software interacts with the 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 simplicity, a "proxy for intelligence" unrelated to compression. If hypotheses are evaluated in terms of weakness, rather than length, we are able to make objective claims regarding performance. Subsequently, we propose objectively optimal notions of AGI and ASI such that the former is computable and the latter anytime computable (though impractical).

Keywords: enactivism · dualism · artificial general intelligence.

1 Introduction

AIXI [1] provides us with a mathematically precise notion of AGI. Its performance is measured according to Legg-Hutter intelligence [2], a proxy for "the ability to satisfy goals in a wide range of environments" [3]. It employs Solomonoff Induction [4, 5] 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) [6]. We explore this problem, and formulate an approach that combines enactive cognition [7], pancomputationalism [8] and weakness as a proxy for intelligence [9].

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 more philosophical than mathematical in nature (a companion to this paper [9] focuses on the mathematical aspects of this research). As such, the following explanation of AIXI is informal and involves some abuse of notation.

Models: A model can be understood as a program [1, 10] or set of rules [11, 12] 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 [13]. 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. To satisfy goals, AIXI must predict[1] 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 [13].

Universal priors: How AIXI obtains an accurate representation of the world can be informally understood in two parts[2]. 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 are most likely to predict the future most accurately. 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. The KC of that model 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 used to measure simplicity and so identify explanations that

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

[2] Again, must be emphasised that this explanation is very informal - the point is just to provide some context to explain the problem of subjectivity.

are more likely to be true. AIXI prefers models that have smaller KC, and in doing so maximises the accuracy of its predictions[3]. AIXI estimates one thing (model accuracy), by measuring another seemingly unrelated thing (KC). In other words, it uses compression as a proxy to estimate intelligence. This proxy for intelligence gives AIXI what is called "a universal prior" [4, 5], 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 (informally) in two steps:

1. Discard any models which "predict" a different the past from the one that actually happened.
2. Use a proxy for intelligence (compression, or more specifically Kolmogorov Complexity) to decide which among the remaining models will most accurately predict the future.

1.2 Subjectivity

KC is measured in the context of a UTM [6]. 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 [6]. To explain in simplified terms (with some abuse of notation), assume a program f_1 is software, f_2 is an interpreter and f_3 is the reality (physical body and environment) within which goals are pursued. Intelligence is a measure of the performance of $f_3(f_2(f_1))$. AIXI is the optimal choice of f_1 to maximise the performance of $f_3(f_2(f_1))$. However, in an interactive setting the perception of success may not match reality.

> "Legg-Hutter intelligence [2] is measured with respect to a fixed UTM. AIXI is the most intelligent policy if it uses the same UTM." [6, 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. $f_3(f_2(f_1))$ depends upon $f_2(f_1)$, not f_1 alone. Thus the performance of f_1 alone is considered to be *subjective*.

> "This undermines all existing optimality properties for AIXI." [6, p.1]

[3] 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).

A UTM is an interpreter. As Leike and Hutter pointed out, Legg-Hutter intelligence is measured with respect to a fixed interpreter, and the problem disappears if AIXI uses that same interpreter. The problem is that there is no way to know what the correct interpreter is. 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 reality, and reality as software. Using the informal notation 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 still depends on the hardware interprets it, we just tend to standardise system architectures. An alternative to dualism is enactivism [7] 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 [8]. Pancomputationalism holds that everything is a computational system. It follows that we may regard the interpreter f_2 as the universe, and reality 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 reality f_3. We do so by merging agent and environment into a task [13], in a sense formalising snapshots of Heidegger's Dasein (being-in-the-world and bound by context) [21].

2.1 A model of reality within reality

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 reality 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 reality. 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 (states of reality). *A set H, where:*

- *We assume a set Φ whose elements we call* ***states****, one of which we single out as the* ***present state*** *of reality*[4]*).*
- *A* ***declarative program*** *is a function $f : \Phi \rightarrow \{true, false\}$, and we write P for the set of all programs. By* ***objective truth*** *about a state ϕ, we mean a declarative program f such that $f(\phi) = true$.*
- *Given a state $\phi \in \Phi$, the* ***objective totality*** *of ϕ is the set of all objective truths $h_\phi = \{f \in P : f(\phi) = true\}$.*
- *$H = \{h_\phi : \phi \in \Phi\}$*

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

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 [11]. 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, "the ability to satisfy goals in a wide range of environments" suggests goals are represented separately from the model of the environment. This too is an unnecessary complication. As Hubert Dreyfus pointed out, creating stored representations for everything is a mistake. After all

> "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. It is the instantiation of intent - a snapshot of Heidegger's being, bound by context [21]. Because of this we will refer to "the mechanism" instead of "the agent" going forward. 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 reality 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. This finite circuitry is a language, albeit one whose meanings are implemented in the pancomputationalist's universe rather than interpreted by a human mind. This language will then be used to formally describe tasks.

---

[4] 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 of reality. For example two different states of reality may be reality from the perspective of two different points in time, or in space and so on.

Definition 2 (implementable language). *A triple $\mathcal{L} = \langle H, V, L \rangle$, where:*

- *H is reality, the set containing all **objective totalities**.*
- *$V \subset \bigcup_{h \in H} h$ is a finite set, named the **vocabulary**.*
- *$L = \{l \in 2^V : \exists h \in H \ (l \subseteq h)\}$, the elements of which are **statements**.*

(Truth) *If we have a statement $l \in L$ expressed using an implementable language, and the totality of the present state of reality is $h \in H$, then l is **true** if $l \subset h$.*
(Extensions) *The **extension of a statement** $a \in L$ is $Z_a = \{b \in L : a \subseteq b\}$, while the **extension of a set of statements** $A \subseteq L$ is $Z_A = \bigcup_{a \in A} Z_a$.*[5]

The programs in V are the circuitry with which cognition is enacted. Only programs included in V can directly impact decision making. We assume cognition always takes place in the context of a physical machine or sensorimotor system – an implementable language. Subsequently, the reader can now safely assume mathematical symbols refer to members or subsets of L unless indicated otherwise. With these, we can define a task.

Definition 3 (task). *A task*[6] *is a triple $\mathcal{T} = \langle S, D, M \rangle$ where:*

- *$S \subset L$ is a set of statements called **situations**, where Z_S is the set of all possible **decisions** which can be made in those situations.*
- *$D \subset Z_S$ is the set of **correct decisions** for this task.*
- *$M \subset L$ is the set of all valid **models** (models) for the task, where*

$$M = \{v \in L : Z_S \cap Z_v \equiv D, \forall z \in Z_v \ (z \subseteq \bigcup_{d \in D} d)\}$$

(How a task is completed) The mechanism is:

1. *presented with a situation $s \in S$, then*
2. *selects $z \in Z_s$, called a decision.*
3. *If $z \in D$, then the agent has made a correct decision and the task will be completed.*

From an ostensive definition of a task (a subset of S and D) a hypothesis $\mathbf{h}$ can be inferred, and a decision $z \in Z_s \cap Z_{\mathbf{h}}$ selected. If $\mathbf{h} \in M$, then $z \in D$.

A single decision instead of sequential decisions: Where AIXI deals in sequential decisions over time [1], a task is completed with a single decision. There are several reasons for this:

[5] A lower case letter is a statement, and upper case a set of statements. The capital letter Z with a subscript indicates the extension of whatever is in the subscript. For example the extension of a statement a is Z_a, and of a set of statements A is Z_A.

[6] For example, this could represent chess as a supervised learning problem where $s \in S$ is the state of a chessboard, $z \in Z_s$ is a sequence of moves by two players that begins in s, and $d \in D$ is a sequence of moves that resulted in victory for one player in particular (represented by the task).

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 [25]. Not all tasks involve sequences of decisions, but all must involve at least one. If a single decision will suffice, why complicate matters by representing sequences?
2. A single decision may set in motion continuous interactions. The preference for sequences may have been a result of reinforcement learners using discrete, pre-defined actions (reasoning about symbolic abstractions). However, in the enactive context such abstractions are not given (thoughts and actions are both just sensorimotor activity).
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.

Binary correctness: To further simplify matters, correctness is binary. Given a task, a decision is considered to be either correct or incorrect. 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 [26]. 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, as well as relative grades thereof, are beyond this paper's scope. Preferences are formalised in a companion to this paper [27, 28].

Representing the past to predict the future: Earlier we described how an accurate model can be obtained in a two step process:

1. Discard any models which "predict" a different the past from the one that actually happened.
2. Use a proxy for intelligence to determine which among the remaining models most accurately predict the future.

In the context of a task, with isolated (as opposed to sequential) decisions that are either correct or not, the past can be represented as the set D of decisions which were deemed correct, along with the situations S in which they were made. A model $m \in M$ entails D given S in that $Z_m \cap Z_S = D$, but may imply a decision $d \notin D$ if presented with a situation $s \notin S$ for which no correct decision is known. In other words, the models in M are equivalent with respect to the past but may disagree about the future situations, addressing step 1 above.

3 The objectively optimal hypothesis

Having formulated cognition so that both the agent and environment are software, 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, the vocabulary of an implementable language is finite (optimal performance can be attained via rote memorisation), and 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 [29] as it is arguably a special case of Kolmogorov Complexity based on slightly different assumptions (compressing data written in a language in an archive written in that same language), however selecting hypotheses by length would still render any claim regarding performance subjective. Third, we must now show that not only is an optimal hypothesis objectively so given a finite vocabulary, but define the objectively optimal choice of vocabulary. To address the former two we require a measure of performance, and an alternative proxy for intelligence, which are addressed [13, 9] by companion to this paper concerning optimal hypotheses.

3.1 Performance

Informally, we define performance as the ability to generalise from limited information, the justification for which is addressed elsewhere [11, 9, 10].

Definition 4 (generalisation). *Given two tasks $\mathcal{T}_\alpha = \langle S_\alpha, D_\alpha, M_\alpha \rangle$ and $\mathcal{T}_\omega = \langle S_\omega, D_\omega, M_\omega \rangle$, a model $m \in M_\alpha$ generalises to task $\mathcal{T}_\omega$ if $m \in M_\omega$.*

Tasks are assumed to be uniformly distributed, and the probability of a statement $l \in L$ generalising to a randomly chosen task $\langle S, D, M \rangle$ written in a language $\langle H, V, L \rangle$ is $p(l \in M \mid l \in L) = \frac{2^{|Z_l|}}{2^{|L|}}$. Optimal performance is attained when the probability of generalisation is maximised. $p(l \in M \mid l \in L)$ is maximised when $l \in \underset{l \in L}{\arg\max}\ |Z_l|$. Assume there exists a ground truth task $\omega = \langle S_\omega, D_\omega, M_\omega \rangle$ written in $\langle H, V, L \rangle$ to which we wish to generalise. The mechanism's knowledge is represented by a task $\alpha = \langle S_\alpha, D_\alpha, M_\alpha \rangle$ such that $S_\alpha \subset S_\omega, D_\alpha \subset D_\omega$ and $M_\alpha \cap M_\omega \neq \emptyset$[7]. The mechanism selects a hypothesis m. The performance of the mechanism is a consequence of $p(m \in M_\omega \mid m \in M_\alpha)$. Optimal performance is attained by $m \in \underset{m \in M_\alpha}{\arg\max}\ p(m \in M_\omega \mid m \in M_\alpha)$.

3.2 Weakness as a proxy for intelligence

First, we must explain why description length is an unsuitable proxy. The description length of m is the cardinality $|m|$ of m itself. For every conceivable task α there exists a program $u \in P$ such that $Z_{\{u\}} = D_\alpha$. If $u \in V$ 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 [9]. Any claim regarding the performance of a mechanism using length as a proxy would be subjective. Optimal performance, in the context of a language $\langle H, V, L \rangle$, is given by $\underset{m \in M_\alpha}{\arg\max}\ p(m \in M_\omega \mid m \in M_\alpha) = \underset{m \in M_\alpha}{\arg\max}\ |Z_m|$ (the relevant proofs, as well as experimental evidence and code supporting those proofs, can be found

[7] In the absence of knowledge, $|Z_l|$ is maximised when $l = \emptyset$.

in [13, 9]). The cardinality of the extension $|Z_m|$ is called the "weakness" of m. Regardless of the language employed, the weakest model remains the optimal hypothesis. This is because the weakest m is necessary and sufficient [9, prop. 1, 2] to maximise $\frac{2^{|Z_m|}}{2^{|L|}}$. $m \in \underset{m \in M_\alpha}{\arg\max}\ |Z_m|$ is the objectively optimal hypothesis, in the sense that it is optimal given any choice of language or task.

3.3 Objectively optimal AGI and ASI

If the cited proof [9] that the weakest hypothesis is optimal holds, then we would suggest well defined notions of AGI and ASI. An AGI is an agent that selects the optimal hypothesis for any given task. An ASI is an AGI that selects the optimal vocabulary to maximise the utility of intelligence for that task. If $\mathbf{h}$ represents our AGI mechanism's hypothesis then

$$\mathbf{h} \in \underset{m \in M_\alpha}{\arg\max}\ |Z_m|$$

given knowledge α. The weakest model is computable via search [13], but such an approach is computationally complex. For an ASI we need to represent a task independently of an implementable language. The powerset 2^P is the set of all possible vocabularies that can be used to specify an implementable language. Let T_V be the set of all tasks that can be written in a vocabulary V, and $T_\forall = \bigcup_{V \in 2^P} T_V$ be the union of all such sets. Let λ be a function $\lambda : 2^P \rightarrow T_\forall$ that takes a vocabulary and returns a particular task in that vocabulary (Λ the set of all such functions). λ represents a task independently of any one implementable language. Assume λ is the task for which we need the ASI. The quantity $\underset{m \in M}{\arg\max}(|Z_m| - |D|)$ expresses the utility of a intelligence given a task [13]. If $\epsilon : \Lambda \rightarrow \mathbb{N}$ is a function such that $\epsilon(\langle S, D, M \rangle) = \underset{m \in M}{\arg\max}(|Z_m| - |D|)$, then an ASI is a mechanism that searches through possible vocabularies to find V such that the utility of intelligence is maximised, and then builds an AGI with that vocabulary. More formally, if $\mathbf{h}$ represents our ASI mechanism's hypothesis, then

$$V \in \underset{V \in 2^P}{\arg\max}\ \epsilon\left(\lambda(V)\right) \text{ and } \lambda(V) = \langle S, D, M \rangle \ \Rightarrow\ \mathbf{h} \in \underset{m \in M}{\arg\max}\ |Z_m|$$

An anytime computable alternative is $V \in \underset{V \in K}{\arg\max}\ \epsilon(\lambda(V))$, where $K \subseteq 2^P$ is the set of vocabularies for which ϵ has been computed so far.

References

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

[2] S. Legg and M. Hutter. “Universal Intelligence: A Definition of Machine Intelligence”. In: *Minds and Machines* 17.4 (2007), pp. 391–444.
[3] S. Legg. “Machine Super Intelligence”. PhD thesis. Uni. of Lugano, 2008.
[4] R. Solomonoff. “A formal theory of inductive inference. Part I”. In: *Information and Control* 7.1 (1964), pp. 1–22.
[5] R. Solomonoff. “A formal theory of inductive inference. Part II”. In: *Information and Control* 7.2 (1964), pp. 224–254.
[6] 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.
[7] D. Ward, D. Silverman, and M. Villalobos. “Introduction: The Varieties of Enactivism”. In: *Topoi* 36 (Apr. 2017).
[8] G. Piccinini and C. Maley. “Computation in Physical Systems”. In: *The Stanford Encyclopedia of Philosophy*. Sum. 21. Stanford University, 2021.
[9] M. T. Bennett. “The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest”. In: *Manuscript* (2023).
[10] F. Chollet. *On the Measure of Intelligence*. 2019.
[11] M. T. Bennett. “Symbol Emergence and the Solutions to Any Task”. In: *Artificial General Intelligence*. Ed. by B. Goertzel, M. Iklé, and A. Potapov. Cham: Springer, 2022, pp. 30–40.
[12] 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.
[13] M. T. Bennett. “Computable Artificial General Intelligence”. In: *Manuscript* (May 2022).
[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*. Ed. by B. Goertzel, M. Iklé, and A. Potapov. Cham: Springer, 2022, pp. 41–44.
[25] H. Kautz and B. Selman. "Planning as satisfiability". In: *IN ECAI-92*. New York: Wiley, 1992, pp. 359–363.
[26] F. M. Artinger, G. Gigerenzer, and P. Jacobs. "Satisficing: Integrating Two Traditions". In: *Journal of Economic Literature* 60.2 (2022), pp. 598–635.
[27] M. T. Bennett. "Emergent Causality & the Foundation of Consciousness". In: *Manuscript* (2023).
[28] M. T. Bennett. "How to Compute Meaning & Lovecraftian Horrors". In: *Manuscript* (2023).
[29] J. Rissanen. "Modeling By Shortest Data Description*". In: *Autom.* 14 (1978), pp. 465–471.