Is Complexity an Illusion?

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

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

Abstract. Simplicity is held by many to be the key to general intelligence. Simpler models tend to "generalise", identifying the cause or generator of data with greater sample efficiency. The implications of the correlation between simplicity and generalisation extend far beyond computer science, addressing questions of physics and even biology. Yet simplicity is a property of form, while generalisation is of function. In interactive settings, any correlation between the two depends on interpretation. In theory there could be no correlation and yet in practice, there is. Previous theoretical work showed generalisation to be a consequence of "weak" constraints implied by function, not form. Experiments demonstrated choosing weak constraints over simple forms yielded a $110 - 500\%$ improvement in generalisation rate. Here we show that all constraints can take equally simple forms, regardless of weakness. However if forms are spatially extended, then function is represented using a finite subset of forms. If function is represented using a finite subset of forms, then we can force a correlation between simplicity and generalisation by making weak constraints take simple forms. If function is determined by a goal directed process that favours versatility (e.g. natural selection), then efficiency demands weak constraints take simple forms. Complexity has no causal influence on generalisation, but appears to due to confounding.

Keywords: complexity · weakness · causality · AGI · information theory.

1 Introduction

Complexity is a quality of systems we find difficult to understand. Formal analogues include entropy [1], compression [2, 3] and even fractal dimension [4]. Physicist Leonard Susskind believes complexity may be the key to a unified theory of physics [5, 6]. Cyberneticist Francis Heylighen recently argued that goals are attractors of dynamical systems that self organise in complex reaction networks [7]. If complex reaction networks do self organise as he argues [8], then it goes some way towards explaining the origins of goal directed behaviour, and thus life. Finally, many hold that complexity is the key to general intelligence [9]. Language models like GPT-4 amount to compressed representations of human language [10]. Simpler objects can be compressed to greater extents, because they exhibit self similarity. As a result, some hold that compression is general intelligence, meaning a general reinforcement learning agent like AIXI [11] can use Solomonoff Induction [12] to maximise expected reward across a wide range of environments. Yet for all that complexity seems to be at the heart of every matter, it has profound flaws. As a qualitative indicator of how *subjectively* difficult a system might be to understand, it makes perfect sense. It makes far less as an indicator of anything *objective*. Ockham's Razor is the epistemic principle that simpler statements are more likely to hold true. It can be understood as the claim that our subjective perceptions of complexity reflect an objective property of our environment. There is no obvious reason this should be the case, and yet it is [13]. Simpler statements tend to be more accurate representations of reality. The aforementioned Solomonoff Induction formalises Ockham's Razor, meaning AIXI is based on the premise that *simpler* models are more accurate depictions of the environment than complex models of seemingly equal predictive power. AIXI is a superintelligence in the sense that it maximises Legg-Hutter intelligence [9]. However, Jan Leike later showed that any claim regarding AIXI's performance is "entirely subjective" [14]. Legg-Hutter intelligence is measured with respect to a fixed Universal Turing Machine (UTM), and AIXI is only optimal if it uses exactly the same UTM. This calls into question the viability of complexity based induction systems in interactive settings. Their performance is subjective, and from a pragmatic standpoint it is only objective performance claims that matter. Leike's result suggests there could be no correlation between objective performance and subjective complexity. This concurs with what seems intuitively obvious, that complexity is an aspect of interpretation. Complexity is a measure of *form*, not *function*. So why does the subjective perception of simplicity tend to correlate with objective performance?

What exactly is complexity supposed to indicate?: As it is used in AGI research [11, 15], complexity is intended to help us infer the program which generated or *caused* data. If one can identify that which caused past data, then one can "generalise" to predict the outcomes in future interactions, to maximise performance [16]. We are concerned with adaptation or "the ability to generalise" [17], not any specific circumstance. This is because any system can eventually identify cause given enough data and memory (by simply rote learning every outcome until it has a complete behavioural specification of the causal program). So assuming one *can* correctly infer cause, then we claim that the amount of data one requires to do so is the sole measure of performance[1]. We refer to this as sample efficiency. The more sample efficiently one can infer cause, the greater one's ability to generalise and adapt to any desired end. Thus, we take intelligence to be a measure of the sample efficiency in generalisation.

Key questions: We build upon previous work [18, 16], in which:

---

[1] Using the same dataset, not different datasets which could necessarily imply a different set of sufficient causes and thus affect learning.

1. Maximising simplicity of policies was proven unnecessary and insufficient to maximise sample efficiency [18, prop. 3].
2. Maximising policy "weakness" was proven necessary and sufficient to maximise sample efficiency [18, prop. 1, 2] and identify cause [16]. In experiments, weak policies outperformed simple by $110 - 500\%$.

Our purpose here is to extend this work, and to establish:

1. Is complexity just an artefact of abstraction?
2. Why do sample efficiency and simplicity tend to be correlated?

Results: We begin by presenting a formalism. Our results are only meaningful if one accepts that our formalism is reflective of reality, so we provide an argument to the effect that it is (lest we be accused of straw-manning complexity). **Second**, we show that the complexity of all behaviours is equal in the absence of an abstraction layer (a general formalisation of any interpreter). In other words, complexity is a subjective "illusion". We further show that if the vocabulary is finite then weakness can confound simplicity and sample efficiency. **Third**, we argue that abstraction is goal directed. If the vocabulary is finite, and tasks uniformly distributed, then weak statements take simple forms.

2 The Formalism

The following definitions are shared with [19, 20], which apply them to biological and philosophical perspectives.

Definition 1 (environment).

- *We assume a set Φ whose elements we call* ***states****.*
- *A* ***declarative program*** *is $f \subseteq \Phi$, and we write P for the set of all declarative programs (the powerset of Φ).*
- *By a* ***truth*** *or* ***fact*** *about a state ϕ, we mean $f \in P$ such that $\phi \in f$.*
- *By an* ***aspect of a state*** *ϕ we mean a set l of facts about ϕ s.t. $\phi \in \bigcap l$. By an* ***aspect of the environment*** *we mean an aspect l of any state, s.t. $\bigcap l \neq \emptyset$. We say an aspect of the environment is* ***realised***[2] *by state ϕ if it is an aspect of ϕ.*

Definition 2 (abstraction layer).

- *We single out a subset $\mathfrak{v} \subseteq P$ which we call* ***the vocabulary*** *of an abstraction layer. If $\mathfrak{v} = P$, then we say that there is no abstraction.*
- *$L_\mathfrak{v} = \{l \subseteq \mathfrak{v} : \bigcap l \neq \emptyset\}$ is a set of aspects in $\mathfrak{v}$. We call $L_\mathfrak{v}$ a formal language, and $l \in L_\mathfrak{v}$ a* ***statement****.*
- *We say a statement is* ***true*** *given a state iff it is an aspect realised by that state.*
- *A* ***completion*** *of a statement x is a statement y which is a superset of x. If y is true, then x is true.*

---

[2] Realised meaning it is made real, or brought into existence.

- *The **extension of a statement**[3] $x \in L_\mathfrak{v}$ is $E_x = \{y \in L_\mathfrak{v} : x \subseteq y\}$. E_x is the set of all completions of x.*
- *The **extension of a set of statements** $X \subseteq L_\mathfrak{v}$ is $E_X = \bigcup_{x \in X} E_x$.*
- *We say x and y are **equivalent** iff $E_x = E_y$.*

(notation) *E with a subscript is the extension of the subscript*[4].

(intuitive summary) *$L_\mathfrak{v}$ is everything which can be realised in this abstraction layer. The extension E_x of a statement x is the set of all statements whose existence implies x, and so it is like a truth table.*

Definition 3 ($\mathfrak{v}$-task). *For a chosen $\mathfrak{v}$, a task α is a pair $\langle I_\alpha, O_\alpha \rangle$ where:*

- *$I_\alpha \subset L_\mathfrak{v}$ is a set whose elements we call **inputs** of α.*
- *$O_\alpha \subset E_{I_\alpha}$ is a set whose elements we call **correct outputs** of α.*

*I_α has the extension E_{I_α} we call **outputs**, and O_α are outputs deemed correct. $\Gamma_\mathfrak{v}$ is the set of **all tasks** given $\mathfrak{v}$.*

(generational hierarchy) *A $\mathfrak{v}$-task α is a **child** of $\mathfrak{v}$-task ω if $I_\alpha \subset I_\omega$ and $O_\alpha \subseteq O_\omega$. This is written as $\alpha \sqsubset \omega$. If $\alpha \sqsubset \omega$ then ω is then a **parent** of α. $\sqsubset$ implies a generational hierarchy of tasks. The level of a task α in this hierarchy is the largest k such there is a sequence $\langle \alpha_0, \alpha_1, ...\alpha_k \rangle$ of k tasks such that $\alpha_0 = \alpha$ and $\alpha_i \sqsubset \alpha_{i+1}$ for all $i \in (0, k)$. A child is "lower level" than its parents*[5].

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

(intuitive summary) *To reiterate and summarise the above:*

- *An **input** is a possibly incomplete description of a world.*
- *An **output** is a completion of an input [def. 2].*
- *A **correct output** is a correct completion of an input.*

Learning and inference definitions: Inference requires a **policy** and learning a policy requires a **proxy**, the definitions of which follow.

Definition 4 (inference).

- *A $\mathfrak{v}$-task **policy** is a statement $\pi \in L_\mathfrak{v}$. It constrains how we complete inputs.*
- *π is a **correct policy** iff the correct outputs O_α of α are exactly the completions π' of π such that π' is also a completion of an input.*

---

[3] The relation to typical philosophical and linguistic notions of intension and extension is addressed at length in [21, 22, 23].

[4] e.g. E_l is the extension of l.

[5] Practical examples child and parent tasks are in a separately published paper with the publicly available experimental code [18].

- *The set of all correct policies for a task α is denoted Π_α.*[6]

Assume $\mathfrak{v}$-task ω and a policy $\pi \in L_\mathfrak{v}$. Inference proceeds as follows:

1. *we are presented with an input $i \in I_\omega$, and*
2. *we must select an output $e \in E_i \cap E_\pi$.*
3. *If $e \in O_\omega$, then e is correct and the task "complete". $\pi \in \Pi_\omega$ implies $e \in O_\omega$, but $e \in O_\omega$ doesn't imply $\pi \in \Pi_\omega$ (an incorrect policy can imply a correct output).*

(intuitive summary) *To reiterate and summarise the above:*

- *A **policy** constrains how we complete inputs.*
- *A **correct policy** is one that constrains us to correct outputs.*

In functionalist terms, a policy is a "causal intermediary".

Definition 5 (learning).

- *A **proxy** $<$ is a binary relation on statements, and the set of all proxies is Q.*
- *$<_w$ is the **weakness** proxy. For statements l_1, l_2 we have $l_1 <_w l_2$ iff $|E_{l_1}| < |E_{l_2}|$.*
- *$<_d$ is the **description length** or **simplicity** proxy. We have $l_1 <_d l_2$ iff $|l_1| > |l_2|$.*

*By the **weakness** of an **extension** we mean its cardinality. By the weakness of a **statement**, we mean the cardinality of its **extension**. Likewise, when we speak of **simplicity** with regards to a **statement**, we mean its cardinality. The complexity of an **extension** is the simplicity of the simplest statement of which it is an extension*[7]*.*

(generalisation) *A statement l **generalises** to a $\mathfrak{v}$-task α iff $l \in \Pi_\alpha$. We speak of **learning** ω from α iff, given a proxy $<$, $\pi \in \Pi_\alpha$ maximises $<$ relative to all other policies in Π_α, and $\pi \in \Pi_\omega$.*

(probability of generalisation) *We assume a uniform distribution over $\Gamma_\mathfrak{v}$. If l_1 and l_2 are policies, we say it is less probable that l_1 generalizes than that l_2 generalizes, written $l_1 <_g l_2$, iff, when a task α is chosen at random from $\Gamma_\mathfrak{v}$ (using a uniform distribution) then the probability that l_1 generalizes to α is less than the probability that l_2 generalizes to α.*

---

[6] To repeat the above definition in set builder notation:

$$\Pi_\alpha = \{\pi \in L_\mathfrak{v} : E_{I_\alpha} \cap E_\pi = O_\alpha\}$$

[7] For example, if we have a language $L_\mathfrak{v}$, and $X \subset L_\mathfrak{v}$ is the set of all statements in $L_\mathfrak{v}$ that all have the extension E_X, then the complexity of E_X is the cardinality of a statement $x \in X$ s.t. there is not statement $y \in X$ with smaller cardinality than x.

(sample efficiency) *Suppose* $\mathfrak{app}$ *is the set of* ***a****ll* ***p****airs of* ***p****olicies. Assume a proxy* $<$ *returns* 1 *iff true, else* 0. *Proxy* $<_a$ *is more sample efficient than* $<_b$ *iff*

$$\left(\sum_{(l_1,l_2)\in\mathfrak{app}} |(l_1 <_g l_2) - (l_1 <_a l_2)| - |(l_1 <_g l_2) - (l_1 <_b l_2)| \right) < 0$$

(optimal proxy) *There is no proxy more sample efficient than* $<_w$, *so we call* $<_w$ *optimal. This formalises the idea that "explanations should be no more specific than necessary" (see Bennett's razor in [18]).*

(intuitive summary) *Learning is an activity undertaken by some manner of intelligent agent, and a task has been "learned" by an agent that knows a correct policy. Humans typically learn from "examples". An example of a task is a correct output and input. A collection of examples is a child task, so "learning" is an attempt to generalise from a child, to one of its parents. The lower level the child from which an agent generalises to parent, the "faster" it learns, the more sample efficient the proxy. The most sample efficient proxy is* $<_w$.

3 Arguments and Results

The distinction between software and hardware is unsuitable for reasoning about cause [16]. The performance of a software agent in an interactive setting is *subjective* [14], as its behaviour depends on hardware which interprets it. Hardware is an "abstraction layer" between software and the surrounding environment. If we are to understand complexity, we must understand what the concept entails in the absence of such abstraction layers. We must ascertain what is *objective* rather than *subjective*, and so we must begin at the level of the environment rather than the agent. To this end, previous work [18, 16, 23] proposed a "de facto" pancomputational [24] model of all conceivable environments and aspects thereof, which we refine and extend. While the formalism used here is a departure from past work, it is equivalent with respect to those previously published proofs we reference [18]. The formalism is not computational in the sense of relying on symbols, quantities, or any other high level abstraction interpreted by a human mind. The assumptions we make are extremely weak (they hold in all conceivable environments).

Axiom 1: When there are things, we call these things the environment[8].

---

[8] It might seem absurd to state something so minimal, but it is necessary to be precise about how minimal our assumptions are.

Axiom 2: The environment has at least one "state"[9]. If there is more than one state, then there is at least one "dimension"[10] a long which things can differ.

A dimension is a set of points, for example time. We do not make the additional claim that such sets of points must be ordered. Each state is the environment at a different point in one or more dimensions. States are "extended" along dimensions, meaning no two states can occupy the same points in all dimensions (in other words, states are like the environment perceived from different positions in time or some other dimensions by an omniscient observer). These are all the assumptions we need for our version of pancomputationalism. We don't even need to speculate about any internal structure states might have, or what dimensions might be. Any "fact" about a state's internal structure can be defined by its relation to other states. The existence of sets is implied by the existence of states (because there is more than one "thing" that exists), and a fact is just the set of states in which it holds (the truth conditions of a "declarative program").

Universality Claim: Axioms 1 and 2 hold for every conceivable environment.

As truth is defined in existential terms[11], there is nothing which is not a fact of the environment, and no environment which does not amount to a set of facts. In other words, this is a minimalist formalism of everything. An environment could be like our own, or not. It could be deterministic, in which case states follow a sequence. It could be non-deterministic, in which case they don't[12]. It could be fantastical with magic and true names. It could even be a world constructed through the subjective experience of its inhabitants. All that matters is that an environment has states, and from the relations between them we obtain the set of all facts. Facts as relations between states let us avoid anything like a universal set, because states are otherwise irreducible. An **aspect** of a state is just a set of facts about that state. With aspects in hand, we can define abstraction. An aspect is akin to a logical **statement**. It has a truth value given a state. We also define its **extension** (def. 2), which is all other aspects of which it is a part. This implies a heirarchy or "lattice" of aspects. Intuitively, an **abstraction layer** is like a window through which one can view

---

[9] Or if the reader prefers, there are as many environments as there are states. By "state" of the environment we mean the aforementioned things. If two states are not the same, then *something* is not the same.

[10] By dimension, again, we just mean something which can differentiate states.

[11] This does not mean "true as interpreted by an omniscient observer", although it does no harm to think of it that way if it helps intuition. For a practical example, the physical state of a transistor in a computer is a declarative program, but so is everything else that might exist. That's the point of deriving this form of pancomputationalism from first principles. We start from "things", and there is more than one state of things when something differs. How this relates to problems of consciousness is addressed elsewhere [16, 23], and is beyond the scope of this paper.

[12] In any case, the difference between deterministic and non-deterministic seems meaningless when you consider that DFAs and NFAs are equivalent [25].

part of the environment. A laptop computer could function as an abstraction layer, as could all or part of the system in which an embodied and embedded organism enacts cognition [26]. In precise terms an abstraction layer is implied by a **vocabulary**, which is a set of declarative programs. A vocabulary implies a formal language whose rules are determined by relations between states. An abstraction layer implies a set of "$\mathfrak{v}$-tasks" (def. 3)[13], each of which is behaviour that defines a system. The formalisation of policies as causal intermediaries between inputs and outputs [29] then develops this into a causal depiction of goal directed behaviour (assuming a policy is implied by the inputs and outputs). For example, an organism could be a policy for a $\mathfrak{v}$-task which is behaviour that organism might enact. Again, we must emphasise this is a first principles approach. We do not assume symbols and Turing machines. Inputs, outputs and policies are all just sets of declarative programs. Whether something is goal directed is determined by the relations between states. The environment makes only one sort of **value** judgement (existence or non-existence), and is otherwise impartial. Goal directed behaviour is a value judgement, so we formalise this impartiality as a **uniform distribution** over tasks. Finally, we must now define complexity. The claims we make in the rest of this paper pertain to this notion of complexity.

Complexity of Extension: The complexity of an extension is the cardinality of the simplest statement of which it is the extension (see def. 5). This is like other formal notions of complexity [2, 3], but facilitates comparison of abstraction layers.

3.1 Implications for Complexity

Proposition 1 (subjectivity). *If there is no abstraction, complexity can always be minimized without improving sample efficiency, regardless of the task.*

Proof: In accord with definition 2, the absence of abstraction means the vocabulary is the set of all declarative programs, meaning $\mathfrak{v} = P$. It follows that for every $l \in L_{\mathfrak{v}}$ there exists $f \in \mathfrak{v}$ such that $\bigcap l = f$. Statements l and $\{f\}$ are equivalent iff $E_l = E_{\{f\}}$, which is exactly the case here because $\bigcap l = f$. [18, prop. 1, 2] shows maximising weakness is necessary and sufficient to maximise the probability of generalisation, which means weakness maximises sample efficiency (is the optimal proxy). This means sample efficiency is determined by the cardinality of extension. For every correct policy l of every task in $\Gamma_{\mathfrak{v}}$ there exists $f \in \mathfrak{v}$ s.t. $E_l = E_{\{f\}}$. Policy complexity can be minimised regardless weakness, because the simplest representation of every extension is simplicity 1. ■

In this sense, complexity is an illusion created by abstraction. In the absence of any particular abstraction, all behaviours (extensions) are implied by statements

---

[13] The notion of task used here descends from the mirror symbol hypothesis [21, 22, 23], however it is complemented by thematically similar research defining tasks in relation to machine learning and biology [27, 28].

of the same complexity. To be clear, we are *not* repeating the claim made by others [14] that *if* the interpreter used by a complexity based induction system matches one used to to compute an objective value for complexity, then that induction system will be optimal in the sense of *eventually* learning the correct policy[14]. We are claiming that if interpretation is truly objective, then $\mathfrak{v} = P$ and complexity has nothing to do with intelligence[15]. There *is* no objective notion of complexity. However, when we take empirical measurements it is inevitably through an abstraction layer, for which $\mathfrak{v} \neq P$. In that context simpler forms have been observed to generalise more efficiently. This raises the question; what additional assumptions can we make that would explain the correlation?

Time, space and causal confounding: We now make the additional assumption that vocabularies are finite. Every aspect of the world in which we exist appears to be spatially extended, meaning no two things occupy the same space at the same time. For the sake of understanding complexity we assume this is true of all environments. We hold that this justifies the assumption of a finite vocabulary, because in our spatially extended environment the amount of information in a bounded system is finite [30].

Proposition 2 (confounding). *If the vocabulary is finite, then policy weakness can confound*[16] *sample efficiency with policy simplicity.*

Proof: We already have that policy weakness causes sample efficiency, in that it is necessary and sufficient to maximise it in order to maximise sample efficiency. Continuing from proof 1, in a finite vocabulary, there may not exist $f \in \mathfrak{v}$ s.t. $E_l = E_{\{f\}}$, which means the complexity of all extensions will not be the same. If we choose any vocabulary in which weaker aspects take simpler forms, then simplicity will be correlated with weakness and so will also be correlated with sample efficiency. This means we would choose $\mathfrak{v}$ s.t. for all $a, b \in L_{\mathfrak{v}}$, the simpler statement has the larger extension, meaning $a <_w b \leftrightarrow a <_d b$. For example, suppose $P = \{a, b, c \dots\},\ a = \{1, 2, 4\},\ b = \{1, 3, 4\},\ \mathfrak{v} = \{a, b\},\ L_{\mathfrak{v}} = \{\{a\}, \{b\}, \{a, b\}\}$, then it follows $\{a, b\} <_w \{a\},\ \{a, b\} <_w \{b\},\ \{a, b\} <_d \{a\},\ \{a, b\} <_d \{b\}$. ■

Why confounding tends to occur: We now briefly argue that abstraction is goal directed. This means the tasks an abstraction layer tends to represent are those it is best suited to represent, which implies weak constraints take simple forms. There are several reasons an abstraction layer is biased toward particular goals, depending upon the context in which we consider complexity. In the case of a computer, a human has specifically designed each abstraction layer to express that which is needed for a purpose. What separates x86 from a higher level

---

[14] To quote verbatim: "Legg-Hutter intelligence is measured with respect to a fixed UTM. AIXI is the most intelligent policy if it uses the same UTM." [14]

[15] Intelligence here meaning not just *eventual* generalisation, but the *efficiency* thereof.

[16] A *confounds* B and C when for example A = "`badly injured`" *causes* B = "`died`" and C = "`picked up by ambulance`", and it looks like C causes B because $p(B \mid C) > p(B \mid \neg C)$, and yet it may be that $p(B \mid C, A) < p(B \mid \neg C, A)$.

abstraction layer like Numpy is that the former has a more general intended purpose, expressing "weaker" constraints. We tend to construct abstraction layers to be as versatile as possible whilst satisfying a particular need. More generally natural selection favours adaptation, which means generalisation, which is maximised by preferring weaker policies [18]. Biological cognition is not limited to the brain [31], meaning the mind is not neatly confined within a well defined neurological abstraction layer. Instead the multiscale competency architectures observed in living organisms [32] amount to self organising abstraction layers. Because natural selection favours adaptable organisms, these abstraction layers will be selected to represent the *weakest* policies which constitute fit behaviour (a weaker policy is more adaptable). More generally, we speculate that phase transitions motivate the emergence of self preserving goal directed behaviour, by destroying some physical structures and preserving others. Such goal directed abstraction *must* minimise the size of vocabularies at higher levels, whilst also maximising the weakness of the policies they can express (two opposing pressures). This is because a larger a vocabulary exponentially increases the space of outputs and policies [22], which may conflict with finite time and space constraints. A larger vocabulary would make inference and learning less tractable (more "complex" in the sense of being a more difficult search problem that takes up more time). To maximise the weakness of policies in higher levels of abstraction, while minimising the size of the vocabulary in which they're expressed, weaker policies must take simpler forms.

References

[1] O. Maroney. "Information Processing and Thermodynamic Entropy". In: *The Stanford Encyclopedia of Philosophy.* 2009.

[2] A. Kolmogorov. "On tables of random numbers". In: *Sankhya: The Indian Journal of Statistics* A (1963), pp. 369–376.

[3] J. Rissanen. "Modeling By Shortest Data Description*". In: *Autom.* 14 (1978), pp. 465–471.

[4] M. F. Barnsley. *Fractals Everywhere.* 2nd ed. Academic Press, 1993.

[5] A. Gefter. "Theoretical physics: Complexity on the horizon". In: *Nature* 509.7502 (May 2014), pp. 552–553.

[6] L. Susskind. "Computational Complexity and Black Hole Horizons". In: *Fortsch. Phys.* 64 (2016), pp. 24–43.

[7] F. Heylighen. "The meaning and origin of goal-directedness: a dynamical systems perspective". In: *BJLS* 139.4 (June 2022), pp. 370–387.

[8] F. Heylighen. "Complexity and Self-organization". In: *Encyclopedia of Library and Information Sciences* (Jan. 2008).

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

[10] G. Deletang et al. "Language Modeling Is Compression". In: *The Twelfth International Conference on Learning Representations.* 2024.

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

[12] R. Solomonoff. "Complexity-based induction systems: Comparisons and convergence theorems". In: *IEEE TIT* 24.4 (1978), pp. 422–432.

[13] E. Sober. *Ockham's Razors: A User's Manual.* Cambridge Uni. Press, 2015.

[14] J. Leike and M. Hutter. "Bad Universal Priors and Notions of Optimality". In: *Proceedings of The 28th COLT, PMLR* (2015), pp. 1244–1259.

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

[16] M. T. Bennett. "Emergent Causality and the Foundation of Consciousness". In: *Artificial General Intelligence.* Springer, 2023, pp. 52–61.

[17] F. Chollet. *On the Measure of Intelligence.* 2019.

[18] M. T. Bennett. "The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest". In: *Artificial General Intelligence.* Springer, 2023, pp. 42–51.

[19] M. T. Bennett. *Meat Meets Machine! Multiscale Competency Enables Causal Learning.* Under review. 2024.

[20] M. T. Bennett. "Computational Dualism and Objective Superintelligence". In: *Artificial General Intelligence.* Springer, 2024.

[21] M. T. Bennett and Y. Maruyama. "Philosophical Specification of Empathetic Ethical Artificial Intelligence". In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2021), pp. 292–300.

[22] M. T. Bennett. "Symbol Emergence and the Solutions to Any Task". In: *Artificial General Intelligence.* Cham: Springer, 2022, pp. 30–40.

[23] M. T. Bennett. "On the Computation of Meaning, Language Models and Incomprehensible Horrors". In: *Artificial General Intelligence.* Springer, 2023, pp. 32–41.

[24] G. Piccinini. *Physical Computation: A Mechanistic Account.* Oxford University Press, June 2015.

[25] M. O. Rabin and D. Scott. "Finite Automata and Their Decision Problems". In: *IBM Journal of Research and Development* 3.2 (1959).

[26] E. Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind.* Cambridge MA: Harvard University Press, 2007.

[27] L. M. Eberding, A. Sheikhlar, and K. R. Thórisson. "SAGE: Task-Environment Platform for Autonomy and Generality Evaluation". In.

[28] R. Cao and D. Yamins. "Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle". In: *Cognitive Systems Research* 85 (2024), p. 101200.

[29] H. Putnam. "Psychological Predicates". In: *Art, mind, and religion.* Uni. of Pittsburgh Press, 1967, pp. 37–48.

[30] J. D. Bekenstein. "Universal upper bound on the entropy-to-energy ratio for bounded systems". In: *Phys. Rev. D* 23 (2 Jan. 1981), pp. 287–298.

[31] A. Ciaunica, E. V. Shmeleva, and M. Levin. "The brain is not mental! coupling neuronal and immune cellular processing in human organisms". In: *Frontiers in Integrative Neuroscience* 17 (2023).

[32] P. McMillen and M. Levin. “Collective intelligence: A unifying concept for integrating biology across scales and substrates”. In: *Communications Biology* 7.1 (Mar. 2024), p. 378.